

Evolutionary applications of population genetics with a focus on malaria: a computational approach

Thesis submitted in accordance with the requirements of the University of Liverpool
for the degree of Doctor in Philosophy by

Tiago Rodrigues Antão

October 2011

Declaration

In compliance with the University of Liverpool rules on PhD thesis, I hereby present my contributions to all the chapters (the majority of which represent published papers). I was the major author of all the manuscripts presented in this thesis save the Mcheza manuscript, produced in joint-first authorship with Mark Beaumont. I also co-authored (as a minor author) another manuscript, *Biopython: freely available Python tools for computational molecular biology and bioinformatics* (Cock et al., 2009) where I collaborated by creating the population genetics code; this manuscript is not presented in this thesis.

For each chapter, my contribution was as follows:

- **ogaraK: A population genetics simulator for malaria** I designed and implemented all the code. The original model is based on Hastings (2006) with extensions from Antao and Hastings (2011b) added by the author. Support for mutation and artesunate models (all drugs are assumed to share one gene in regards to drug resistance) is also of my responsibility.
- **Environmental, pharmacological and genetic influences on the spread of drug resistant malaria** The original model of this publication was defined in Hastings (2006). I extended it with: epistasis, multiple treated “environments” (i.e. semi-immune hosts versus no immunity or full treatment compliance versus partial treatment compliance) and multiple drugs.
- **The promise and dangers of recent antimalarial deployment policies**
Single author.
- **Evolutionary parasitology applied to control and elimination policies**
Single author.
- **LOSITAN: A workbench to detect molecular adaptation based on a F_{st} -outlier method** *BMC Bioinformatics* requires information about the contributions of each author. The information supplied is provided here, copied verbatim: “TA is the leading architect and main developer of LOSITAN, and drafted this

publication. ABP and GL have both theoretically drafted the idea of developing LOSITAN and together with TA, RJL contributed in discussions, planning and writing of this manuscript. RJL developed the web page and tutorials and AL developed the code regarding multi-core detection and graphics and data export.” Perhaps more important than all other co-authors is the work of Mark Beaumont in the application FDIST which sits at the core of LOSITAN (see the paragraph below). Indeed, LOSITAN is never cited alone (and should be never cited alone), as it is always cited alongside with Beaumont and Nichols (1996).

- **Mcheza: A workbench to detect selection using dominant markers** The only chapter produced in joint-first authorship. Mark Beaumont supplied the DFDIST application. I supplied an altered version of the LOSITAN code.
- **interPopula: a Python API to access the HapMap Project dataset** Single author.
- **Detecting F_{st} -outliers and selection requires genotyping multiple SNPs per gene: lessons from empirical data** Gordon Luikart supplied the analysis related to next generation sequencing (i.e. the subsection titled “SNP genotyping and discovery”). All the other content was mine.
- **Assessing selection for drug resistant malaria: Can temporal F_{ST} help?** I conceived most of the study, Gordon Luikart suggested doing simulations of neutral loci, which I conducted. Ian Hastings assured that the models implemented reflected realistic malaria scenarios.
- **Early detection of population declines: High power of genetic monitoring using effective population size estimators** Both Andrés Pérez-Figueroa and Gordon Luikart suggested the parameters for the models. I designed the experimental procedure, implemented all the code, run the simulations and performed most of the analysis.
- **Estimating effective population size of disease vectors: a critical assessment of applications and performance** Martin Donnelly did the literature search on vectors and suggested the parameters for the simulations. Gordon Luikart was involved in the evaluation of the LD method. Ian Hastings gave input on issues related to transmission control. I designed all the study (with special mention to the novel sinusoidal model of population size), implemented all the code and did the vast majority of the analysis.
- **Interpreting estimates of effective population size in parasites and vectors** Andrés Pérez-Figueroa and Gordon Luikart provided input about long term N_e estimation. Gordon Luikart also provided input about heterozygosity changes.

Ian Hastings provided input on linkage disequilibrium in parasites and Martin Donnelly commented on temporal methods applied to disease vectors. I conceived and structured this chapter as an opinion piece and designed and implemented all the simulations.

I also note that, for all manuscripts above, I was responsible for writing the vast majority of the text (though co-authors supplied some corrections and valuable input).

The work presented in this thesis has not been submitted for any other degree.

Tiago Rodrigues Antão

Abstract

Malaria is a major public health concern for the one-third of the human population estimated to be exposed to the threat of the most virulent species, *Plasmodium falciparum*. Modern molecular and computational tools from population genetics may help to better understand and fight the burden of drug resistant malaria.

Malaria biology is substantially different from the underlying paradigm of standard population genetics models, most notably because malaria has both a asexual haploid phase and a sexual diploid phase where selfing (i.e. mating between genetically identical parasites) is possible. It is therefore fundamental to understand if commonly used population genetics methods are robust to the deviations from standard expectations imposed by the malaria life-cycle.

We build novel models of malaria population genetics and provide guidelines to interpret empirical studies of the spread of drug resistance. Using realistic models of epistasis between genes involved in drug resistance we suggest that all signals of linkage disequilibrium (LD) are possible and that researchers should be confident in reporting a lack of statistical association between genes involved in resistance to antimalarials. We also suggest that researchers should be cautious in interpreting changes in the prevalence of drug resistance after control interventions as reductions in transmission can cause a change in prevalence without concomitant change in frequency of resistance.

We provide guidelines to better design and interpret studies related to estimating effective population size (N_e). We use computational simulations to study scenarios that can approximate control and elimination interventions (i.e. where a significant part of the population is killed). For N_e estimation our results suggest that researchers must increase the number of individuals and loci genotyped in order to have sufficiently precise N_e estimates. LD-based N_e estimators are more appropriate for early detection of control and elimination interventions than temporal-based N_e estimators. Long-term estimators based on heterozygosity should not be used to make inferences about contemporary demographic processes. We also applied our analysis to disease vectors and we concluded that LD-based estimation is able to detect demographic seasonality

patterns (i.e. changes in population size due to variations imposed by wet and dry seasons) whereas temporal estimators will provide averages over longer periods of time.

We also studied selection detection using F_{ST} -outlier approaches. Our results suggest that temporal F_{ST} might not be appropriate for early detection of genes involved in drug resistance (e.g. in the case of artesunate derivatives). We also provide software and guidelines to better design and interpret studies (also across other taxa) of selection based on F_{ST} -outlier approaches. Most notably our results suggest that sampling only one or two SNPs per locus (as it is done in many empirical studies) might not be sufficient to detect areas of the genome under selection, and that at least 4 SNPs per loci should be genotyped.

Contents

Declaration	i
Abstract	v
Contents	vii
List of publications	xi
List of Tables	xiii
List of Figures	xv
Abbreviations	xvii
1 Introduction	1
1.1 Population genetics models of drug resistant <i>P. falciparum</i>	4
1.2 F_{ST} -outlier selection detection and discovering genes involved in drug resistance	6
1.3 Effective population size and assessing the success of control and elimination measures	7
1.4 Acknowledgements	9
I Population genetics models of drug resistant malaria	11
2 ogaraK: A population genetics simulator for malaria	13
2.1 Introduction	13
2.2 Approach	14
2.3 Discussion	15
3 Environmental, pharmacological and genetic influences on the spread of drug resistant malaria	17

3.1	Model and methods	20
3.2	Results	22
3.3	Discussion	27
4	The promise and dangers of recent antimalarial deployment policies	33
4.1	Modeling and theory	36
4.2	Results	39
4.3	Discussion	44
4.4	Conclusion	46
5	Evolutionary parasitology applied to control and elimination policies	53
II	F_{ST} selection detection and discovering genes involved in drug resistance	55
6	LOSITAN: A workbench to detect molecular adaptation based on a F_{ST}-outlier method	57
6.1	Background	57
6.2	Implementation	58
6.3	Results and Discussion	61
6.4	Conclusions	62
6.5	Availability and requirements	63
7	Mcheza: A workbench to detect selection using dominant markers	65
7.1	Introduction	65
7.2	Software implementation	66
7.3	Discussion	67
8	interPopula: a Python API to access the HapMap Project dataset	69
8.1	Background	69
8.2	Implementation	70
8.3	Results	72
8.4	Conclusions	74
8.5	Availability and requirements	74
9	Detecting F_{st}-outliers and selection requires genotyping multiple SNPs per gene: lessons from empirical data	77
9.1	Introduction	77
9.2	Methods	79
9.3	Results	80
9.4	Discussion	84

9.5	Conclusions and recommendations	88
10	Assessing selection for drug resistant malaria: Can temporal F_{ST} help?	91
10.1	Introduction	91
10.2	Methods	93
10.3	Results	96
10.4	Discussion	100
10.5	Conclusion	103
III	Estimating effective population size and assessing the success of control and elimination measures	105
11	Early detection of population declines: High power of genetic monitoring using effective population size estimators	107
11.1	Introduction	107
11.2	Methods	108
11.3	Results	112
11.4	Discussion	116
12	Estimating effective population size of disease vectors: a critical assessment of applications and performance	125
12.1	Introduction	126
12.2	Methods	129
12.3	Results	132
12.4	Discussion	136
12.5	Conclusion	140
13	Interpreting estimates of effective population size in parasites and vectors	143
13.1	Introduction	143
13.2	Problems with heterozygosity based N_e estimation	145
13.3	Contemporary estimation of N_e	147
13.4	Conclusions and guidelines	147
IV	Discussion and conclusion	151
14	Discussion and conclusion	153
14.1	Population genetics models of drug resistant <i>P. falciparum</i>	155
14.2	F_{ST} selection detection and discovering genes involved in drug resistance	156

14.3 Estimating effective population size and assessing the success of control and elimination measures	157
14.4 Final remarks	159
Bibliography	161
Appendices	177
A ogaraK: Supplementary material	179
A.1 Simulator overview	179
A.2 Drug policies	185
A.3 Simple user guide	188
A.4 Software issues	192
A.5 Example analysis	193
B A formal description of the population genetics model to study the spread of drug resistant malaria	197
C Expected heterozygosity: Illustrative examples of a slow moving statistic	201
C.1 Methods	201
C.2 Results and remarks	203

List of publications

The majority of chapters in this thesis describe content already published in scientific journals:

Chapter 2 Antao, T and Hastings, IM. ogaraK: a population genetics simulator for malaria. *Bioinformatics*, 27(9):1335, 2011

Chapter 3 Antao, T and Hastings, IM. Environmental, pharmacological and genetic influences on the spread of drug-resistant malaria. *Proceedings of the Royal Society B: Biological Sciences*, 278(1712):1705–1712, 2011

Chapter 5 Antao, T. Evolutionary parasitology applied to control and elimination policies. *Trends in Parasitology*, 27(6):233–234

Chapter 6 Antao, T, Lopes, A, Lopes, R, et al. LOSITAN: a workbench to detect molecular adaptation based on a F_{ST} -outlier method. *BMC Bioinformatics*, 9(1):323, 2008

Chapter 7 Antao, T and Beaumont, M. Mcheza: A workbench to detect selection using dominant markers. *Bioinformatics*, 27 (12): 1717-1718, 2011

Chapter 8 Antao, T. interPopula: a Python API to access the HapMap Project dataset. *BMC Bioinformatics*, 11(Suppl 12):S10, 2010

Chapter 11 Antao, T, Perez-Figueroa, A, and Luikart, G. Early detection of population declines: high power of genetic monitoring using effective population size estimators. *Evolutionary Applications*, 4(1):144–154, 2011

Chapter 12 is currently submitted to Molecular Ecology. Chapters 9, 10 and 13 will be submitted soon. There is currently no intention to submit chapter 4.

List of Tables

3.1	The outcome of human drug treatment according to genetic modes of resistance	20
3.2	Percentage of scenarios which stabilise at intermediate frequencies of resistance	26
4.1	The impact of key factors in the spread of drug resistance	39
4.2	Impact of MFT, sequential application and combination therapy on useful therapeutic life	39
9.1	Characteristics of the candidate and high- F_{ST} genes studied	81
9.2	Summary information for all genes used	82
9.3	Detection of high F_{ST} SNPs in the studied genes	83
12.1	Representative sample of empirical studies of contemporary N_e in disease vectors	128
A.1	The outcome of drug treatment according to resistance	182
C.1	Mean heterozygosity and effective population size estimated after a bottle- neck or expansion	205

List of Figures

3.1	The frequency of drug sensitive parasites and its rate of change	23
3.2	Linkage disequilibrium patterns for single-drug models	24
3.3	Relationship between prevalence and frequency of resistance with MOI . . .	26
3.4	Prevalence as a function of the frequency of the resistant allele	27
4.1	Impact of policy and MOI on useful therapeutic life with two drugs	41
4.2	The distributions of genotypes at the end-of-life for MFT, sequential appli- cation and combination therapies	42
4.3	The distributions of genotypes at the end-of-life for MFT, sequential appli- cation and combination therapies (end of life is extended to 50% prevalence)	43
4.4	Frequency and linkage disequilibrium (r) with varying policies and epistasis modes	48
4.5	Impact of policy and epistasis on useful therapeutic life with two drugs . .	49
4.6	Impact of policy and MOI on useful therapeutic life with three drugs . . .	50
4.7	Impact of policy and epistasis on useful therapeutic life with three drugs . .	51
6.1	LOSITAN console	59
6.2	Graphical output of LOSITAN	60
8.1	F_{ST} for Lactase	73
8.2	Example code to print the frequency of HapMap SNPs	76
9.1	The F_{ST} distribution for all SNPs in the 5 genes under directional selection and 5 genes in chromosome 2 with the highest proportion of SNPs with F_{ST} above 0.45.	84
9.2	F_{ST} for 1% of SNPs along chromosome 2 and zooming in on four genes sampled from Yorubans in Africa and Utahans representing North Western Europeans	89
9.3	F_{ST} for random haplotype reconstructions using increasing amounts of SNPs	90

10.1	Summary statistics for simulations using three epistasis modes	97
10.2	The ratio between frequency of resistance and its prevalence as a function of the multiplicity of infection	98
10.3	The relationship between prevalence, epistasis mode, multiplicity of infection and temporal F_{ST}	99
10.4	The upper confidence interval of temporal F_{ST} for neutral markers	100
11.1	Power to detect that N_e is below 150% post-bottleneck N_c	113
11.2	Boxplot charts of LD and temporal \hat{N}_e after a bottleneck (I)	114
11.3	Boxplot charts of LD and temporal \hat{N}_e after a bottleneck (II)	115
11.4	Harmonic mean of \hat{N}_e and 95% CIs (post-bottleneck)	122
11.5	Boxplot of the distribution of \hat{N}_e under equilibrium	123
11.6	Boxplot of the \hat{N}_e using SNPs	124
12.1	Boxplot charts of temporal point estimates obtained from moments-based F method varying time span and sampling strategy	133
12.2	Harmonic mean of \hat{N}_E and 95% confidence intervals of all replicates for the moment-based temporal estimator for a time span between 1 and 24 generations	134
12.3	Boxplot charts for the point estimates of all three N_e estimation methods using different sampling strategies	135
12.4	The behaviour of all estimators with a seasonal model with fluctuating population size	141
12.5	Harmonic mean of \hat{N}_E for all methods for two bottleneck scenarios	142
A.1	Single drug deployment scenario	182
A.2	Single drug deployment scenario: SP mode	183
A.3	Rotation and MFT scenarios	186
A.4	Rotation with mixed mode scenario	187
A.5	The change in frequency of various genomes over time	193
A.6	The impact of deployment policies on temporal F_{ST}	195
C.1	Boxplot charts of the distribution of N_e point estimates for scenarios of constant N_e using different sampling strategies	206
C.2	Bar charts for the estimation of N_e in bottleneck and expansion scenarios .	207
C.3	Mean \hat{N}_e for four founding scenarios	208

Abbreviations

ACT	Artemisinin Combination Therapy
AFLP	Amplified Fragment Length Polymorphism
ANTXR1	Anthrax toxin receptor 1 gene
API	Application Programming Interface
CAD	Carbamoylphosphate synthetase 2/aspartate transcarbamylase, and dihydroorotase gene
CAB39	Calcium binding protein 39 gene
CEU	Utah Residents with ancestry from Northern and Western Europe (HapMap population)
CI	Confidence Interval
CLASP1	Cytoplasmic linker associated protein 1 gene
CNV	Copy Number Variation
CQ	Chloroquine
CRT	Chloroquine Resistance Transporter gene
CV	Coefficient of Variation
DArT	Diversity Arrays Technology
DDT	Dichlorodiphenyltrichloroethane
DGF	Duplicate Gene Function
DHFR	Dihydrofolate reductase gene
DHPS	Dihydropteroate synthase gene

DNA Deoxyribonucleic acid

FDR False Discovery Rate

FE Full Epistasis

FOSL2 Fos-related antigen 2 gene

KITLG Kit ligand gene

IRS Indoor Residual Spraying

ITN Insecticide-Treated bed Nets

LCT Lactase

LD Linkage Disequilibrium

MAP4K3 mitogen-activated protein kinase kinase kinase 3 gene

MDR1 Multidrug Resistance 1 gene

MDM4 Mdm4 p53 binding protein homolog gene

MFT Multiple First-line Therapies

ML Maximum Likelihood

MOI Multiplicity of infection

MRSA Methicillin-resistant *Staphylococcus aureus*

OCA2 Oculocutaneous albinism II gene

RAD Restriction-site Associated DNA

SLC24A5 Solute carrier family 24 member 5 gene

SP Sulfadoxine-Pyremethamine

SQL Structured Query Language

TB Tuberculosis

WHO World Health Organization

YRI Yoruban from Ibadan, Nigeria (HapMap population)

One

Introduction

Malaria is a major public health concern for the one-third of the human population estimated to be exposed to the threat of the most virulent species, *Plasmodium falciparum*, with an estimated number of clinical episodes ranging from 300 to 660 million per annum (Snow et al., 2005). There is no effective vaccine for this species, and infection is controlled by insecticides targeted at vector mosquito species, and treatment by anti-malarial drugs. As might be expected, insecticide and drug resistance rapidly evolved and spread (Olliaro, 2005). Modern molecular and computational tools from population genetics may help to better understand and fight the burden of drug resistant malaria. This promise has gradually been fulfilled and the benefits have already been substantial, for instance modern molecular methods showed that resistance to several antimalarial drugs had, contrary to expectation, a very small number of independent origins (Roper et al., 2004; Wootton et al., 2002).

While there is a urgent need to develop new methods to analyse the ever increasing amount of data, especially with the advent of next generation sequencing, there is also the need to understand how existing population genetics approaches are robust to realistic biological assumptions of *P. falciparum* biology and epidemiology. *P. falciparum* genetics (Tuteja, 2007) differs significantly from standard population genetics models (e.g. standard population genetics models assume diploidy, whereas *P. falciparum* has both a asexual haploid and sexual diploid phase) and epidemiological considerations, for instance control interventions (like treatment with antimalarials or the use of bed nets) potentially impose changes on the size of the population, whereas most population genetics models assume either constant sized or infinite populations. Are existing widely used methods robust to malaria assumptions? How can we improve the reliability of results? Should we simply avoid some approaches? Here we will try to answer some of these questions for methods used for discovery of loci under selection (which can be used, e.g., to find genes important in drug resistance) and early detection of population bottlenecks (important to assess the impact of malaria control and elimination measures).

Here we also address the relationship between theoretical and empirical work. Most modern theoretical work has a focus on prediction, for instance the prediction of the impact of drug deployment policies (Maude et al., 2009; Boni et al., 2008) or vaccination strategies (Smith et al., 2006) which has minor applicability to the interpretation of existing (past-related) empirical data. On the other hand it is not uncommon to find empirical researchers who seem to put little emphasis on the importance of theoretical findings. The approach followed here tries to bridge theoretical and empirical work by assessing the impact of, sometimes unconscious, theoretical assumptions made in empirical research work. One example should clarify this: Existing empirical population genetics studies normally assume that association (measured by Linkage Disequilibrium - LD) between loci involved in drug resistance should be positive, therefore linkage equilibrium between those loci is seen as a “negative result”. Such assumption is based on previous sound theoretical work (Dye and Williams, 1997) but more recent findings about the mode of action between genes involved in drug resistance of Sulfadoxine-Pyrimethamine (SP), where 2 genes are involved in drug resistance but one, *dhfr* is more important than the second *dhps* (Olliaro, 2001) require a revision of the idea of equal importance of loci. This can have, and indeed we will show that it has, impact on the expectation of positive LD.

The questions addressed in this thesis fall in two broad categories:

1. How do realistic models of the spread of drug resistant *P. falciparum* influence widely used population genetics measures (e.g. linkage disequilibrium)? Can more realistic population genetics models of drug resistance provide useful insights to better design drug deployment policies, curbing the spread of drug resistant malaria?
2. Can existing methods to detect selection and change in population size be applicable to *P. falciparum* malaria and its vectors? Can we improve experimental design in order to better estimate the impact of control and elimination policies on parasite genetic variability?

The first set of questions will be addressed by developing novel population genetics models of *P. falciparum* drug resistance. These models improve the field of malaria drug resistance modeling by accounting for epistasis relationships between drug resistance genes, selection heterogeneity (e.g. immunity or treatment compliance) and multiple simultaneous drug deployments. The work done to address these questions is imminently theoretical though great care was put on assuring that these novel models are realistic (i.e. can qualitatively approximate known field/empirical results).

In order to answer the second set of questions we will leverage existing knowledge in the field of conservation genetics. As it will become clear some of the existing methods are developed in the context of conservation (i.e. small population sizes) and it is not

guaranteed that the parasitology community has even absorbed some more basic results of population genetics associated with the use of some of these methods. Indeed, it is fundamental to disseminate some very robust results from population genetics, which, if not correctly applied (and we will demonstrate that incorrect applications do exist) will cause misleading interpretations of field data resulting from control interventions. We will also discuss the implications of using methods and approaches that were mostly developed in the context of management and protection of species (i.e. conservation) versus a context of control, elimination and ultimately, eradication (parasitology). The work done to address these questions has a clear empirical approach: it is geared towards empirical data (re)analysis and provides constructive criticism with regards to commonly use data analysis strategies.

This thesis is split into three parts, one solely dedicated to the first set of questions (i.e. modeling the spread of drug resistance to understand its impact on data analysis and control interventions), and two dedicated to the second set of questions: one supporting and evaluating methods for selection detection (specifically F_{ST} -outlier approaches) and one dedicated to evaluating effective population size (N_e) estimators. For the first two parts all the methods (software) used were published in scientific journals, therefore each software application will have a dedicated chapter. Parts two and three, though they are dedicated to evaluating methods to analyse empirical data suffer from the lack of real *P. falciparum* datasets to analyse, such unfortunate event has had a strong impact on the development of this thesis and will be further discussed in this introduction and in the conclusion. We now present each part of this thesis starting with a small overview on its structure.

This thesis is presented in paper format (i.e. as a series of related papers), the majority of them already published in scientific journals. This imposes some constraints on this thesis structure namely:

1. The size, organisation and content of each chapter/paper varies substantially, mostly because of requirements imposed by journals on size and structure and on content by editors and reviewers. The most extreme example is probably chapter 2 describing ogaraK, the software application implementing our population genetics models of drug resistant malaria: Application notes are limited to two pages on the journal *Bioinformatics* and as such most of the content formally describing the model was made available in the supplement (appendix A in this thesis).
2. Some repetition between manuscripts is unavoidable as standalone papers have to have enough information, especially on methods. This is particularly clear when comparing the supplements of chapters 2 (ogaraK) and chapter 3 (Environmental, pharmacological and genetic influences on the spread of drug resistant malaria) whose supplements (appendixes) are, respectively, A and B.

3. The logical order presented here (method/software chapters precede analysis chapters) does not correspond to publication order. For instance, chapter 2 (method) references Antao and Hastings (2011b) (research), but the research chapter (accepted paper on *Proc Roy Soc B*) only points to the software web page, not to Antao and Hastings (2011a) as the software manuscript was published later.
4. Some of the content was imposed by reviewers (including title changes). This issue is further discussed in the conclusion.
5. Minor changes were made to the papers published, most notably, typos were corrected and references to funding sources were removed.

Several chapters reflect software applications (all of them published in scientific journals). The reader is encouraged to use and test the applications, all of them freely available online. The source code for the applications is also freely available and can be inspected and copied by anyone subject to the GNU public license version 3.

We present now an overview of each part of this thesis.

1.1 Population genetics models of drug resistant *P. falciparum*

This part is sub-divided in two: one chapter describing the software used to make simulations and three chapters with analysis based or inspired on the simulator results.

Software

In order to simulate realistic population genetics models of *P. falciparum* accounting for drug resistance a new simulator will be developed. All known previous studies of *P. falciparum* population genetics never made the software (either full applications or mathematical models developed using computer algebra systems) available. As far as we know only one epidemiological simulator (non population genetics based) of malaria is made publicly available (Smith et al., 2008)¹. One of the objectives of this work was to do the computational part as transparent and replicable as possible, making all used software artifacts available for public inspection and re-use. While the best solution would be to use already existing general-purpose individual forward-time population genetics simulators, these were not suitable to simulate malaria for two main reasons:

1. As individual simulation is impossible from a computational performance perspective (individual-based simulation becomes prohibitive considering the number of different parasites, which can rise to 10^{12} inside a human).

¹During the course of this thesis, the author participated in the development of this other simulator, though no content related to that effort is included here.

2. No existing simulator is capable of simulating a genome that has both a haploid and diploid phase over its life-cycle. The above limitations apply even to the most flexible population genetics simulator, simuPOP (Peng and Kimmel, 2005).

Therefore we will design and implement ogaraK (Antao and Hastings, 2011a) (chapter 2), a population genetics simulator designed for malaria. ogaraK is originally based on the model presented in (Hastings, 2006), extended to support (Antao and Hastings, 2011b) and simulates the frequency of infection genotypes (not individual parasites): it is thus a model of infinite-sized populations (i.e., no drift). It will allow us to simulate, among other factors, the genetic interactions leading to resistance, different human environments (e.g. untreated humans, treated humans with no immunity, treated humans with semi-immunity) and the multiplicity of infection (influencing inbreeding) inside each human. OgaraK can simulate different drug deployment policies like rotation of drugs over time (imposing different selection pressure over time) or multiple-first line therapies (Boni et al., 2008) (imposing different selection pressures to different infections). As a side effect, unrelated to *P. falciparum* biology, and as ogaraK will be able to model flexible epistasis (genetic interactions) and different patterns of selection pressure, it can also be used to simulate the Red-Queen Hypothesis and spacial selection heterogeneity used in sex-theory (Otto, 2009).

Research

Using ogaraK we will investigate in (Antao and Hastings, 2011b) (chapter 3, appendix B) how epistasis, inbreeding, selection heterogeneity and multiple simultaneous drug deployments interact to influence the spread of drug resistant malaria. We will study how different human “environments” within which treatment may occur (such as semi- and non-immune humans taking full or partial drug courses) influence the genetic interactions between parasite loci involved in resistance. We will discuss how the rate of spread varies according to different malaria transmission intensities, why resistance might stabilise at intermediate frequencies and also identify several factors that influence the decline of resistance after a drug is removed. We will try to understand how different transmission intensities might bias the conclusions of studies based on clinical outcomes with regards to the spread of resistant parasites. We will also study the importance of epistasis and LD in understanding the impact of transmission reduction measures on the relationship between prevalence and frequency of resistance. Most unfortunately we suggest that the potentially positive impact of transmission control measures on the spread of resistance might be overestimated. This chapter and supplement will also offer a brief introduction to the topic of modeling the spread of drug resistance malaria along with references to pertinent literature.

Current malaria drug deployment policies recommend using a single first-line therapy for most clinical malaria cases but recent research has suggested that using multiple first-

line therapies (MFT) might yield a better clinical outcome by delaying the emergence and spread of drug resistance. In chapter 4 we will compare different drug deployment policies using several realistic parasite population genetics models of the spread of drug resistant malaria. We will account for differences in modes of genetic interaction, proportion of infected humans treated, immunity, treatment compliance and the impact of transmission intensity on inbreeding. We will try to understand if MFT policies can delay the spread of drug resistance, including multiple resistant genotypes. Our work will stress the importance of compliance with treatment guidelines as incomplete treatment can potentially increase the spread of resistant infections. We will also study how association between loci, measured by linkage disequilibrium, is fundamental to understand the dynamics of spread of resistance and the interactions between genes.

Finally, using the acquired knowledge in the previous three chapters, we will briefly comment (Antao, 2011) (chapter 5) on the importance of evolutionary biology in the context of drug deployment policies. We will discuss existing proposals of using sub-curative antimalarial drug treatment and also the possible negative impact of elimination policies on the spread of drug resistance.

1.2 F_{ST} -outlier selection detection and discovering genes involved in drug resistance

In order to help evaluate the performance of F_{ST} -outlier approaches we will develop two applications to reliably apply the method of Beaumont and Nichols (1996) to both co-dominant and dominant markers. The initial objective was to use these applications with *P. falciparum* data. As such data was not available we then will produce an empirical study using human data and a theoretical study using simulated data created with ogoraK and simuPOP (Peng and Kimmel, 2005). In order to access human data we also developed an application to access human data on the public HapMap project (International HapMap Consortium, 2007).

Software

We will develop two applications implementing the F_{ST} -outlier method in (Beaumont and Nichols, 1996) commonly used to detect loci selection. The applications are, LOSITAN (Antao et al., 2008) (chapter 6) for co-dominant markers and Mcheza (Antao and Beaumont, 2011) (chapter 7) for AFLPs. Our implementations will extend the original ones (FDIST and DFDIST, Mark Beaumont, unpublished) by including an easy to use interface, multitest correction and a more reliable approximation of nuisance parameters.

We will also implement interPopula (Antao, 2010) (chapter 8) a library to access the HapMap dataset (International HapMap Consortium, 2007) which includes millions of SNP polymorphisms from 11 different human populations. The HapMap dataset, by its

size (in terms of number of markers, populations and individuals) and the amount of existing knowledge about the human genome allows to test existing population methods like the F_{ST} selection detection method against empirical data and not just synthetic datasets. As we know many loci in the human species that are under selection in certain populations (e.g. Lactase) it is possible to evaluate a selection method against a species and a genome with a (partially) known selection history.

F_{ST} evaluation

As datasets for *P. falciparum* were not available we will study the behaviour of F_{ST} with one empirical dataset (the public human database of HapMap) and one simulated dataset created with ogaraK and simuPOP.

We will study, in chapter 9 the performance of F_{ST} with single nucleotide polymorphisms (SNPs) which are increasingly used to identify genes under selection. However, researchers often genotype only a few SNPs per gene, so we quantify the sensitivity of using only a few SNPs in a gene to identify high F_{ST} -outlier genes using large empirical data sets from humans in the HapMap project. We will try to understand if F_{ST} can detect genes under selection and how many SNPs are necessary to detect it. We will research the impact of genotyping more than one SNP per gene on the false positive rate. We will discuss the consequences of current genotyping policies (sampling a single SNP per gene is not uncommon for many species) and the potential benefits of using next generation sequencing (i.e. more markers across the genome and also per gene).

As *Plasmodium falciparum* malaria is subject to artificial selection from antimalarial drugs which select for drug resistant parasites, detecting which genes are under selection can provide fundamental insights to help containing the spread and the burden of drug resistant malaria. We will evaluate, using computational simulations, the performance of temporal F_{ST} to reliably detect genes under selection. We will try to understand in what conditions temporal F_{ST} is more powerful (e.g. transmission intensity or time between samples). We will try to understand the scope of applicability of temporal F_{ST} in several relevant epidemiological scenarios ranging from the impact of transmission seasonality (i.e. wet and dry seasons) to scanning for the genes involved in Artemisinin resistance. We will also discuss the appropriate sampling strategies (i.e. the necessary number of individuals to be sampled) to have a reliable estimation of F_{ST} .

1.3 Effective population size and assessing the success of control and elimination measures

In an era where malaria control and elimination measures are being scaled up, it is important to have measures of success in terms of parasite genetic diversity and effective population size. It is indeed fundamental to assess, as soon as possible, if such measures

are having a positive impact in reducing parasite population size and diversity. We will start by evaluating the ability to rapidly detect a population decline using two widely-used contemporary effective population size estimators. As such estimators are usually studied in conservation genetics settings (with low N_e) our first analysis was still done using low N_e (i.e. not applicable to *P. falciparum*) as a baseline for further studies. We will then study N_e estimators in the context of disease vectors: disease vectors may have high N_e and are normally subjected to seasonality patterns (usually high population numbers in wet seasons, low population numbers in dry seasons). Finally we will discuss the usage of N_e estimators in the broad context of parasitology (i.e. considering both parasites and vectors), our discussion, though not novel, will point out several misunderstandings in recent parasitology research regarding N_e . We will now present all researched issues in more detail.

To evaluate the ability to rapidly detect a population decline we will compare in Antao et al. (2011) (chapter 11 with supplemental data available on the CD) a two-sample temporal method (Krimbas and Tsakas, 1971) and a one-sample method based on LD (Hill, 1981; Waples, 2006; Waples and Do, 2008). We will use simulated data representing a wide range of population sizes, sample sizes, and number of loci. For this simulation exercise we will use an existing forward-time individual based population genetics application (Peng and Kimmel, 2005) and we will simulate a small number (below 500) of diploid individuals. We will study how many generations are required to detect a bottleneck after such an event happens. We try to understand how many markers (SNPs and microsatellites) and how many individuals need to be sampled in order to have a reliable estimation of N_e . We will also study the relative importance of sampling more individuals or more loci. We will provide some guidelines regarding the design of studies targeted at monitoring population declines.

We then discuss, in chapter 12, the application of contemporary effective population size estimators to insect disease vectors. N_e in disease vectors can have two distinguishing features from common scenarios considered in most N_e studies: high effective population size (above 500) and seasonal population fluctuations (wet- and dry-season). We will research the impact of high N_e on the precision of estimators and the need to increase sampling sizes to cope with loss of precision due to high N_e . We will also try to understand the impact of seasonality on estimator performance.

We then discuss, on chapter 13 and appendix C the interpretation of effective population size estimators and measures of heterozygosity in the context of recent empirical studies done with *P. falciparum* (e.g Gatei et al., 2010; Anderson et al., 2000a; Iwagami et al., 2009; Susomboon et al., 2008) and disease vectors (e.g. Lehmann et al., 1998; Simard et al., 2000; Pinto et al., 2002), we will discuss the validity of some observations made in such research and try to understand if observations are consequence of control interventions or simply a methodological artifact. We will re-interpret recently published

studies in the light of realistic assumptions about the behaviour of heterozygosity and N_e estimators.

While the software to do the simulations required for this part is not published, its most complex parts are publicly available inside the software package newAge (<http://popgen.eu/soft/newAge>). For this part, we essentially use the standard approach for most research based on computational simulations: we simply describe our methodological approach on each chapter.

1.4 Acknowledgements

Firstly I would like to thank Ian Hastings. Then to Gordon Luikart. And also: Mark Beaumont, Peter Cock, Mary Creegan, Martin Donnelly and Andrés Pérez-Figueroa.

Then, of course, there are personal acknowledgements: you know who you are!

This work was supported by research grant SFRH/BD/30834/2006 from Fundação para a Ciência e Tecnologia, Portugal.

Part I

Population genetics models of drug resistant malaria

Two

ogaraK: A population genetics simulator for malaria

Tiago Antao and Ian M. Hastings

Abstract

Motivation: The evolution of resistance in *Plasmodium falciparum* malaria against most available treatments is a major global health threat. Population genetics approaches are commonly used to model the spread of drug resistance. Due to uncommon features in malaria biology existing forward-time population genetics simulators cannot suitably model *Plasmodium falciparum* malaria.

Results: Here we present ogaraK, a population genetics simulator for modelling the spread of drug resistant malaria. OgaraK is designed to make malaria simulation computationally tractable as it models infections, not individual parasites. OgaraK is also able to model the life cycle of the parasite which includes both haploid and diploid phases and sexual and asexual reproduction. We also allow for the simulation of different inbreeding levels, an important difference between high and low transmission areas and a fundamental factor influencing the outcome of strategies to control or eliminate malaria.

Availability: OgaraK is available as free software (GPL) from the address <http://popgen.eu/soft/ogaraK>.

2.1 Introduction

Malaria is a major public health concern, as one third of the human population is estimated to be exposed to the threat of the most virulent species, *Plasmodium falciparum*. Antimalarial drug resistance has emerged as one of the major challenges facing malaria control. Drug resistance became widespread to most first line therapies and treatment failures are now being observed for their replacements, Artemisinin Combination Therapies (ACTs) (Dondorp et al., 2009). Mathematical and computational models of the spread of drug resistance are an important tool to understand the emergence and spread of drug resistance.

Most mathematical and computational modelling of malaria have been based on epidemiology (e.g. Koella and Antia (2003)) or population genetics (e.g. Hastings (1997)), though complex simulation models have also been developed (Smith et al., 2008). While many forward-time population genetics simulators do exist (e.g. Peng and Kimmel (2005)), they are not suitable to model malaria, therefore all existing computational studies using a population genetics approach are based on applications and scripts developed for each study and not directly subjected to peer review or publicly available.

Standard individual-based forward-time population genetics simulators are not suitable to model *P. falciparum* biology for two main reasons: (i) population size in malaria can rise up to 10^{12} parasites per human host, making it computationally infeasible to simulate all individuals and (ii) *P. falciparum* life cycle includes both haploid and diploid phases and most existing simulators do not allow for the modelling of different genotypic structures over time. In order to address these issues we developed ogaraK, a population genetics simulator designed to study the spread of drug resistance in *P. falciparum* malaria.

2.2 Approach

OgaraK features and limitations are based in *P. falciparum* population biology in the presence of treatment pressure. OgaraK was designed to study the spread of existing drug resistance and it can be used to understand how recognised important factors in malaria epidemiology (e.g. different transmission intensities) influence the spread of resistance and also to compare different drug deployment policies.

Drug treatments can be modelled as a form of selection pressure over the parasite population. OgaraK is able to simulate a wide variety of selection pressures modelling both temporal and spacial selection heterogeneity. Temporal heterogeneity is a proxy for a policy of drug rotation while spacial heterogeneity approximates the use of multiple first line drugs.

A wide range of epistasis modes of loci involved in drug resistance are also supported. While most existing theoretical research assumes that parasites require all mutations (full epistasis) related to a drug in order to resist treatment, ogaraK is able to model other epistasis modes, like duplicate gene function or asymmetry. These modes reflect existing empirical evidence for some drugs (e.g. Chloroquine or SP) where genes vary in their importance to confer resistance (Olliaro, 2005). Multiple epistasis modes are also useful to model poor drug compliance or host immunity: A weaker epistasis mode among drug resistance loci is enough to confer resistance in humans with no acquired immunity or who take an incomplete treatment; in humans who take a full course or have acquired partial immunity against malaria, a stronger epistasis mode is required to resist treatment.

Multiplicity of infection (MOI) has been recognised as one of the factors that differentiate between high and low transmission areas of malaria. MOI affects the spread of resistance, recombination and population inbreeding levels as the mating alternatives on the obligatory sexual phase in the mosquito are limited by the number of genetically different parasites ingested in a blood meal. OgaraK allows the simulation of different inbreeding levels, allowing for a varying MOI across simulations and also, within each simulation, simulating individuals with different MOI.

Individual based simulation is replaced by exhaustive enumeration of all possible combinations of infection types (each infection having a specific genotype). This method was first used in Hastings (2006) and comparative analysis between results derived from epidemiological simulations and this approach show consistent results (Boni et al., 2008).

The main purpose of the simulator is to study the spread of resistance but we also support mutation therefore allowing the study of *de novo* emergence. In the case of most antimalarials (e.g. Chloroquine or SP (Wellems and Plowe, 2001)) mutation is a rare event. Furthermore resistance already exists to most drugs, even Artemisinin based therapies (Dondorp et al., 2009), hence our focus on spread of existing mutations.

In order to study simultaneous usage of multiple ACTs which might share one resistance gene (as they all have an Artemisinin derivative as principal component) we also provide models of multi-drug resistance where part of the resistance mechanism is shared among all drugs. A “standard” model where all drugs involved have unrelated loci is also available.

OgaraK has an easy to use interface which can be run from the web as a Java Webstart application. The code is available and can be also linked as a library or used in batch mode.

Results are exported in a text format (reporting the frequency of all genotypes over time) and also made available in the widely used Genepop (Rousset, 2008) format. Simple scripts for data analysis are provided using Biopython (Cock et al., 2009), but these mainly serve as examples as it is expected that most analysis will be done using standard population genetics packages owing to the ability to export data in Genepop format.

A supplement (appendix A) is included where the model is detailed and where example applications and a user manual are also supplied.

2.3 Discussion

OgaraK is able to easily simulate most existing population genetics models studying the spread of drug resistance in malaria. It is made available as a public framework which can be used to evaluate new models or re-use old ones for new analysis. For instance it was already used and tested to research the impact of epistasis on linkage disequilibrium between loci involved in drug resistance (Antao and Hastings, 2011b) and

can also simulate, from a population genetics perspective, promising drug deployment strategies (Boni et al., 2008).

We focused on modelling how resistance spreads and not *de novo* emergence, given that emergence is a rare event and that it is already widespread to most drugs therefore making the management of existing resistance a major concern. Nonetheless the ogaraK supports mutation, therefore permitting the study of *de novo* emergence.

While ogaraK was developed with malaria modelling in mind, epistasis and spacial and temporal selection patterns have clear parallels with some known models in sex theory (Otto, 2009), as temporal selection heterogeneity is the fundamental concept behind the Red-Queen Hypothesis and spacial selection heterogeneity has also been proposed as one explanation for sex and recombination. OgaraK can therefore be used to easily simulate and test some models relevant to sex theory.

OgaraK can serve as a framework to more easily evaluate drug deployment policies and help enhance the understanding of fundamental variables underlying the spread of drug resistant malaria.

Three

Environmental, pharmacological and genetic influences on the spread of drug resistant malaria

Tiago Antao and Ian M. Hastings

Abstract

Plasmodium falciparum malaria is subject to artificial selection from antimalarial drugs which select for drug resistant parasites. We describe and apply a flexible new approach to investigate how epistasis, inbreeding, selection heterogeneity and multiple simultaneous drug deployments interact to influence the spread of drug resistant malaria. This framework recognises that different human “environments” within which treatment may occur (such as semi- and non-immune humans taking full or partial drug courses) influence the genetic interactions between parasite loci involved in resistance. Our model provides an explanation for how the rate of spread varies according to different malaria transmission intensities, why resistance might stabilise at intermediate frequencies and also identifies several factors that influence the decline of resistance after a drug is removed. Results suggest that studies based on clinical outcomes might overestimate the spread of resistant parasites, especially in high transmission areas. We show that when transmission decreases, prevalence might decrease without a corresponding change in frequency of resistance and that this relationship is heavily influenced by the extent of linkage disequilibrium between loci. This has important consequences on the interpretation of data from areas where control is being successful and suggests that reducing transmission might have less impact on the spread of resistance than previously expected.

Malaria is a major public health concern for the one third of the human population estimated to be exposed to the threat of the most virulent species, *Plasmodium falciparum*, with an estimated number of clinical episodes ranging from 300 to 660 million per annum (Snow et al., 2005). There is no effective vaccine for this species, and infection

is controlled by insecticides targeted at vector mosquito species, and treatment by anti-malarial drugs. As might be expected, insecticide- and drug-resistance rapidly evolved and spread (Olliaro, 2005; Rogers et al., 2009; Dondorp et al., 2009). This manuscript focuses on the dynamics of antimalarial drug resistance although the approach we develop herein may be generalised to other control agents such as insecticides, herbicides and anthelmintics.

Plasmodium parasites are haploid and reproduce asexually in humans. Humans often contain several genetically distinct *P. falciparum* clones acquired from different mosquito bites; the number of clones in a human is called the multiplicity of infection (MOI). MOI is a proxy for transmission intensity as higher transmission intensity increases MOI due to repeated sequential infection (Anderson et al., 2000a). *P. falciparum* parasites undergo an obligate sexual phase in the mosquito before being transmitted back into humans as haploids. Mating between gametes from the same clone (selfing) involves sexual recombination between identical haploid genotypes resulting in clonal reproduction. Mating between different clones (outcrossing) results in genetic re-assortment of *P. falciparum* genes. Mating can only occur between clones transmitted from the same human, hence the rate of outcrossing depends on the MOI. Field estimates reveal that outcrossing is relatively common and can occur in more than 50% of matings (Mzilahowa et al., 2007). It has been postulated that clones within a human compete for resources and transmission and that removal of drug-sensitive clones following treatment allows the surviving resistant clones to garner additional resources and to increase their transmission (Hastings, 1997; Hastings and D'Alessandro, 2000), an effect recently termed "competitive release" (Wargo et al., 2007); the higher the MOI the larger the potential effect of competitive release. MOI is therefore fundamental in the dynamics of the spread of resistance as it increases the rate of sexual recombination (outcrossing) which allows parasites with different resistance profiles to mate and also breaks down the association between alleles encoding drug resistance.

Mathematical models play an important role in understanding the forces driving resistance and in designing policies to minimise the rates at which resistance arises and spreads (e.g. Boni et al. (2008); Yeung et al. (2004)). Their importance arises for three main reasons: Firstly, it is near-impossible to address this issue empirically as anti-malarial drug deployments occurs on country – and even continent-wide scales – so the effects of local differences in deployment strategies are likely to be swamped by immigration of resistance driven by national deployment policies (Anderson and Roper, 2005). Secondly, it is difficult to generalise the lessons learnt from individual drugs because their dynamics are likely to differ substantially depending on the genetic basis of resistance and the degree of resistance they encode: for example resistance arises incredibly rarely to Chloroquine (CQ) and Sulfadoxine-Pyrimethamine (SP) (Anderson and Roper, 2005), the two drugs so far deployed worldwide, but very easily to other drugs such as Atovaquone (Looareesuwan et al., 1996) and Pyrimethamine (Plowe et al., 1997), and

at intermediate rates – apparently due to gene duplications rather than point mutations – to Mefloquine (Price et al., 2004). In addition, high-level resistance to Atovaquone occurs in a single mutational step, while resistance to SP requires sequentially accumulation of mutations at several codons in two genes (Sibley et al., 2001). Thirdly, basic population genetics predicts that the alleles encoding resistance increase in frequency exponentially so the most important dynamics occur when resistance is at undetectably low frequencies meaning there are few empirical data on the process.

All models make simplifying assumptions but the primary one we relax and investigate here, is the assumption that infections bearing a “resistant” genotype always survive drug treatment. Empirical evidence shows that the fate of a “resistant” infection, survival or death, is much more probabilistic and depends critically on the human “environment” within which drug treatment and selection occurs. Human immunity plays a huge role and drugs may be highly effective in semi-immune adults but highly ineffective in non-immune infants (Langhorne et al., 2008; Rogerson et al., 2010). Similarly, “sensitive” infections may survive treatment in humans where drug levels are sub-optimal either through poor compliance with the drug regimen or because their pharmacogenetics means that drugs are poorly absorbed or rapidly eliminated (Guerin et al., 2002), while “resistant” infections may be eradicated in humans with high drug levels. It is trivial to incorporate these effects into a single locus model (we simply assign a probability of survival in a treated individual) but dynamics become more realistic and complex in situations where two (or more) loci are required to encode resistance.

When more than one locus is involved in drug resistance it is important to quantify the association between resistance alleles at different loci. Linkage Disequilibrium (LD), the non-random association of resistance alleles at different loci, has been shown to be a critical factor influencing the rate of spread of resistance (Dye and Williams, 1997). Many genetic models have ignored LD and those that have measured it assumed complete epistasis between the mutations (i.e. only infections with mutations at all resistance loci would survive treatment) with the result that LD was always positive between mutations. We allow three different models of parasite genetic interaction: “full epistasis” where, as before, resistant mutation at both loci are required for the infection to survive; “asymmetric epistasis” where one locus has more impact on survival so resistance is determined primarily by that locus irrespective of the allele at the second locus; “duplicate gene action” where a resistant mutation at either locus will allow survival. It is axiomatic among geneticists that the mode of gene action is not fixed but depends on the environment in which they are expressed and this is what we address here: infections encountering “strong” selection environments (humans with high levels of immunity and/or drug) may require full epistasis to survive, while as the environment becomes “weaker” (less immunity and/or drug) then asymmetric and eventually duplicate gene action will best describe the fate of “resistant” mutations. These modes of selection are summarised on Table 3.1.

Resistance profile		Epistasis mode		
Locus 1	Locus 2	Full	Asymmetric	DGF
Sens	Sens	Cure	Cure	Cure
Sens	Res	Cure	Cure	Resistance
Res	Sens	Cure	Resistance	Resistance
Res	Res	Resistance	Resistance	Resistance

Table 3.1: The outcome of human drug treatment according to genetic modes of resistance. Cure occurs if all loci are sensitive. A single mutation at either loci is enough to confer resistance with duplicate gene function (DGF). In asymmetric epistasis, the first locus is necessary and sufficient to confer resistance. Mutations at both loci are necessary with full epistasis. Different environments can be present simultaneously and determine the mode of action: weaker epistasis modes (DGF or asymmetry) might be applicable for individuals with poor drug compliance or who are non-immune, while immune individuals or high drug dosage might require full epistasis. Asymmetric epistasis is only relevant in multi-environment model, where another epistasis mode is also present.

Herein we present a deterministic population genetics model of the spread of drug resistant *P. falciparum* using computational simulations to investigate how these key factors affect the spread of drug resistance (i.e. selection heterogeneity, number of resistance loci) within the context of local malaria epidemiology and local drug policies (for example, whether different drugs are co-deployed to reduce selection for resistance or whether they are rotated such that one is used until it fails and then replaced). The model presented is sufficiently general to study many different parameters but we will present results for scenarios of most practical importance and realism for *P. falciparum* population biology and drug deployment strategies. Practical policy considerations are discussed as we try to understand the implications of control and elimination measures for the spread of resistance.

3.1 Model and methods

We make the following assumptions in line with most previous modelling efforts: that clones co-infecting the same human are genetically unrelated; that clones have equal infectivity and mate at random so that if there are n clones in a human then selfing rate is $\frac{1}{n}$ and outcrossing rate $1 - \frac{1}{n}$, and that competitive release occurs. The spread of drug resistance is tracked using a time-scale of parasite generations (a generation is a parasite reproduction cycle from host to host which is likely to be around 5 per year). Loci are assumed to be physically unlinked, as is the case for loci known to be involved in malaria drug resistance (Osman et al., 2007), and each locus can have two alleles: resistant and sensitive. A resistant allele will incur a fitness penalty in the absence of

the drug and all mutations are assumed to have the same fitness penalty. Genotypes with multiple mutations suffer a multiplicative fitness penalty.

Development of drug resistance can be seen as a two-step process, the *de novo* emergence of the resistant mutation and its subsequent spread. Existing research shows that, for most drugs, resistance emerged extremely infrequently (Wellems and Plowe, 2001), the notable exceptions being Atovaquone (Vaidya and Mather, 2000) and Pyrimethamine (Price et al., 2004; Roper et al., 2004). Understanding the appearance of *de novo* mutations is an important topic, discussed elsewhere (e.g. White and Pongtavornpinyo (2003); Pongtavornpinyo et al. (2009)), so we assume that resistance alleles already exist at very low frequencies at the onset of the simulation and focus on understanding their subsequent spread.

The model is designed to study the spread of drug resistant alleles subjected to different drug deployment policies and a mathematical formalisation of the model is provided in the supplement (appendix B). There are evidently a very large number of parameter combinations that can be explored. Here we concentrate on seven illustrative scenarios, named and described below.

Single locus A single drug is deployed with resistance encoded at a single locus. There are two human environments: treated and untreated individuals. Resistant parasites survive in both environments (but may pay a fitness penalty in untreated humans) while sensitive survive in untreated humans but are cleared in treated. This is obviously the simplest case, explored elsewhere, but is included as a baseline simulation.

Full epistasis A single drug is deployed with resistance encoded by two loci. Two environments are present, untreated and treated, and survival in the treated humans requires resistance at both loci (i.e. full epistasis).

Duplicate gene function (DGF) A single drug is deployed with resistance encoded by two loci. Two environments are present, untreated and treated, with selection in the latter sufficiently weak that resistance alleles at either locus can encode survival.

Asymmetric epistasis A single drug is deployed where resistance is encoded by two loci. Two environments are present, untreated and treated. The first locus is necessary and sufficient to encode resistance in treated individuals and the second is irrelevant. This scenario is therefore functionally identical to the single locus scenario, but asymmetric epistasis serves as a useful model of resistance when used in more complex and realistic environments, in conjunction with full epistasis. Asymmetric epistasis mimics SP resistance where the *dhps* resistance allele involved in the *de novo* folate production pathway cannot fully replace the *dhfr*

gene involved in exogenous folate usage. This model is also applicable for the asymmetric importance of supplementary *mdr* gene to *crt* in Chloroquine based resistance (Olliaro, 2001).

Full epistasis + DGF A single drug is deployed with resistance encoded by two loci. There are three environments: untreated, well-treated (treatment is sufficiently effective that full epistasis is required for survival) and poorly-treated (treatment is sub-optimal so that resistance alleles at either locus can encode survival). The DGF environment can model non-immune individuals, incomplete treatment courses or any other situations where the parasite does not need all mutations to survive. The full epistasis environment models semi-immune individuals, complete treatment courses or other events which require the parasite to have both mutations in order to resist treatment.

SP-based (Full and asymmetric epistasis) A single drug is deployed with resistance encoded by two loci. There are three environments: untreated, well-treated (treatment is sufficiently effective that full epistasis is required for survival) and poorly-treated (treatment is sub-optimal so that resistance alleles at the more important locus can encode survival). The name of this scenario comes from the drug SP where resistance may be encoded by alleles at *dhfr* alone or, in well-treated individuals, may require resistant alleles at both *dhfr* and *dhps* loci.

Two drugs Two drugs are deployed separately (i.e., the parasites never encounter both simultaneously in the same generation) with 2 loci for each drug. There are five environments: untreated, well-treated with drug 1, well-treated with drug 2, poorly-treated with drug 1, poorly-treated with drug 2. Full epistasis is required for survival in well-treated individuals while asymmetrical epistasis determines survival in poorly-treated ones.

In scenarios with more than one treatment environment (i.e. the last three scenarios) it is assumed that there are equal proportions of each environment among the treated infections.

3.2 Results

The following outcomes are possible for each locus in a simulation: (i) one allele (sensitive or resistant) tends towards fixation, or (ii) allele frequencies stabilise at intermediate levels. We opt to describe the dynamics of the fully sensitive form (i.e. with no mutations) in order to simplify the presentation of results.

Figure 3.1 plots the frequency of sensitive alleles together with their rate of change under the one-locus scenario. The rate of change is dependent on the MOI. During

the initial phases of spread, higher MOI entails a faster spread of resistant profiles, but this process is reversed when resistant parasites approach fixation. This reveals that the speed of spread is frequency dependent and consequently, that the total time to resistant fixation is a bad proxy of the speed of spread at low frequencies. This occurs because the total time in most cases is more influenced by what happens at high resistance frequencies rather than the dynamics at lower frequencies; it is the latter part of the dynamics, when resistance is starting to spread and cause drug treatment failures, that have the most implications for drug policy choice. Understanding the spread of resistance at important frequencies therefore requires analysis of the whole behaviour of the model and not just the time until the final outcome.

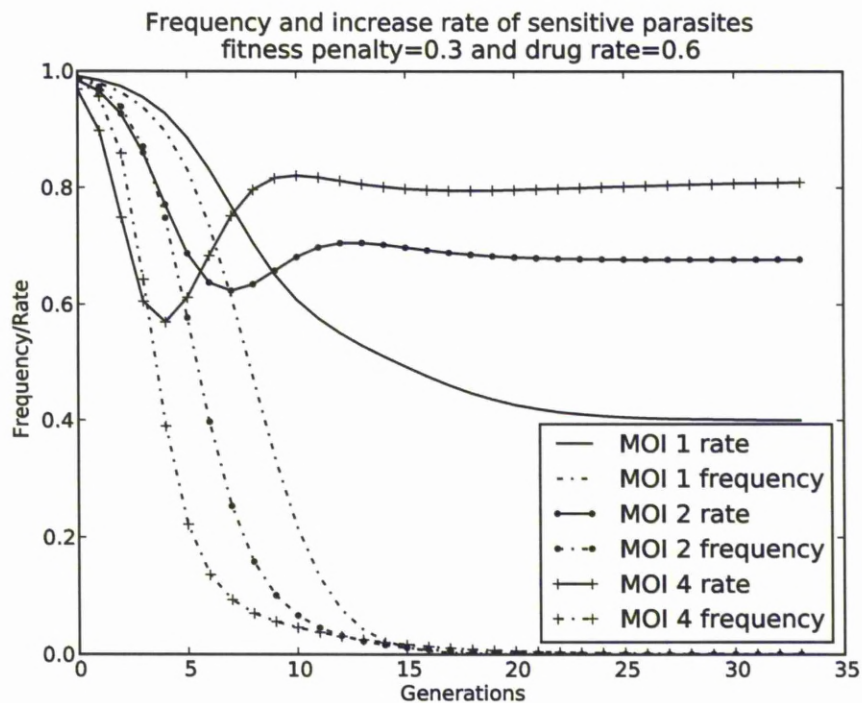


Figure 3.1: The frequency of sensitive parasites and its rate of change computed as the proportionate change per generation. The frequency of sensitive genotypes is depicted in dashed lines and the rate of change in solid lines. The rate is below one as resistance is increasing. At low levels of frequency (which are important for policy decision) bigger MOIs entail a faster decline of sensitive forms. Single locus model shown.

Figure 3.2 shows linkage disequilibrium, r , for the four scenarios where a single drug is deployed and resistance is encoded by two loci. Where full epistasis is required, r is positive (in line with comparable results in Dye and Williams (1997) and Hastings (2006)). This arises because genotypes $c_{0,1}$ and $c_{1,0}$ (a description of the genotype

notation is presented in the supplement) provide no advantage in any environments as they are not resistant to treatment and are less fit than the sensitive clones in untreated individuals. With duplicate gene function, one mutation is sufficient to confer resistance and furthermore two simultaneous mutations are never advantageous, and r becomes negative. When more than one treated environment is available, as the full-epistasis plus DGF scenario and in the SP-based scenario where asymmetrical epistasis plus DGF occur, the opposing effects tend to cancel each other out and LD is low. From a qualitative point of view epistasis has a tremendous impact in both signal and amplitude of LD.

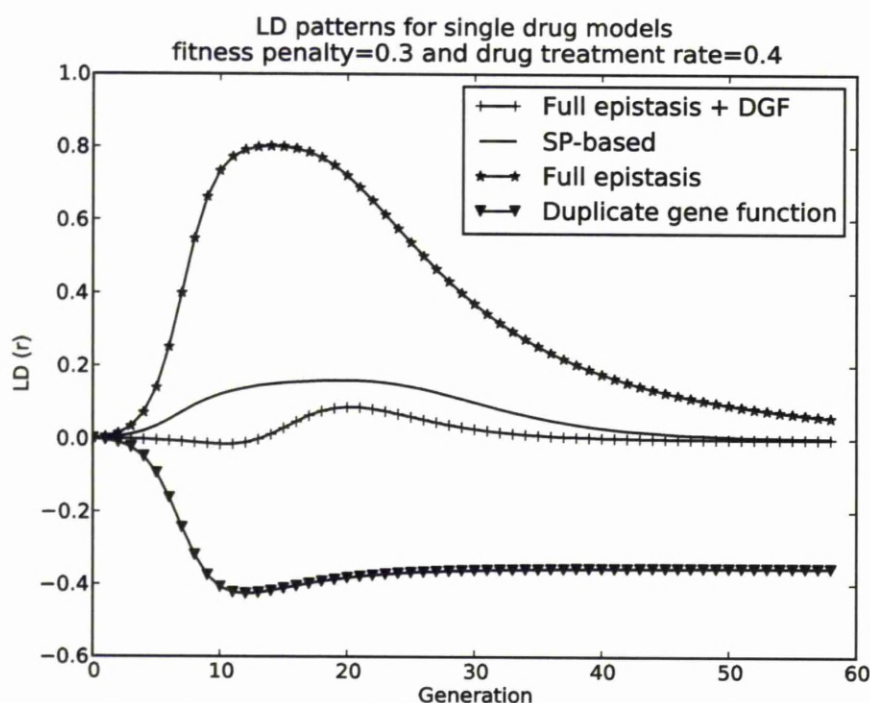


Figure 3.2: Linkage disequilibrium (r) patterns for single-drug models with MOI of 2, drug usage rate of 40% and a fitness penalty of 30%. Different epistasis models have qualitatively different disequilibrium patterns.

Figures 3.3 and 3.4 illustrate how MOI and LD affect the relationship between the prevalence of a resistant mutation and its underlying allele frequency. The prevalence at a single locus, P , is given by the binomial distribution: $1 - (1 - F_1)^i$ and is plotted on Figure 3.3 for various MOI. A bigger MOI implies a bigger prevalence for the same frequency. The difference between frequency and prevalence is not maximised at the extremes frequencies (i.e. near 0.0 or 1.0) but at intermediate frequencies. We note that prevalence is inevitably higher than frequency unless the MOI is 1. This has implications

for the interpretation of field data and is discussed below. It is quite easy to compute the frequency of resistance for the single locus model given the prevalence and the MOI, for more realistic models with two or more loci, the frequency also depends on linkage disequilibrium which is itself dependent on assumptions about epistasis. We can repeat the analysis of the relationship between frequency and prevalence considering one drug with two resistant loci. The prevalence is given by $1 - (1 - F_{1,1})^i$ for full epistasis (this is because only $c_{1,1}$ clones are resistant) and $1 - (1 - (F_{1,1} + F_{1,0} + F_{0,1}))^i$ for DGF ($c_{0,1}$, $c_{1,0}$ and $c_{1,1}$ are all resistant). Assuming equal frequencies for both resistant alleles (a realistic assumption for all epistasis models except asymmetry) the frequency of resistant infections can be easily calculated for both models as a function of frequency of mutation of one gene and the linkage disequilibrium measure r . The relation between frequency of a locus involved in resistance and prevalence as a function of MOI and r is shown on figure 3.4 for an r of 0.0 and 0.5 for full epistasis. As expected higher r entails a bigger proportion of individuals harbouring a resistant infection in a full epistasis scenario (the converse is expected and observed for duplicate gene function). In all cases, higher MOI implies a larger proportion of individuals harbouring one resistant infection, but the behaviour of the function is difficult to quantify with precision.

MOI and epistasis influence whether stabilisation of resistance occurs at intermediate frequencies. Table 3.2 presents the proportion of scenarios which stabilise at intermediate frequencies for the scenarios considered (the supplementary material provides details on the parameter ranges investigated). Single locus and full epistasis models are included for comparison as they have been widely studied before (e.g., Dye and Williams (1997); Hastings (2006)). Our results are consistent with those studies as both scenarios tend to fixation in a very large part of the search space. Duplicate gene function, by making all parasites with two mutations less fit than parasites with just a single mutation even in treated individuals, has a much bigger portion of the search space where stability occurs at intermediate frequencies. This is analogous to the situation of over-dominance in diploids which is known to produce stable allele frequencies. In this model, two non-mutated loci are bad for the parasite (it is drug sensitive), one mutated locus is optimal (resistance) while two mutated loci pay a double fitness penalty: they are the least competitive genotype in untreated hosts and less fit than single-mutated parasites in treated hosts. In all models, increasing MOI increases the size of the parameter space where stable intermediate frequencies occurs.

We also investigated a single-locus scenario where the frequency of resistance was started at 99% and no drugs were used. This allows us to isolate and investigate the effect of untreated individuals on the spread of drug resistance and provides insight into the likely effect of removing a drug from circulation. In the simple situation of MOI=1 then the frequency of resistant infections remains unchanged even if a drug is removed, because the fitness penalty was modelled as competition within the human host (Equation 1 of the Supplement) this assumption was made for convenience but an

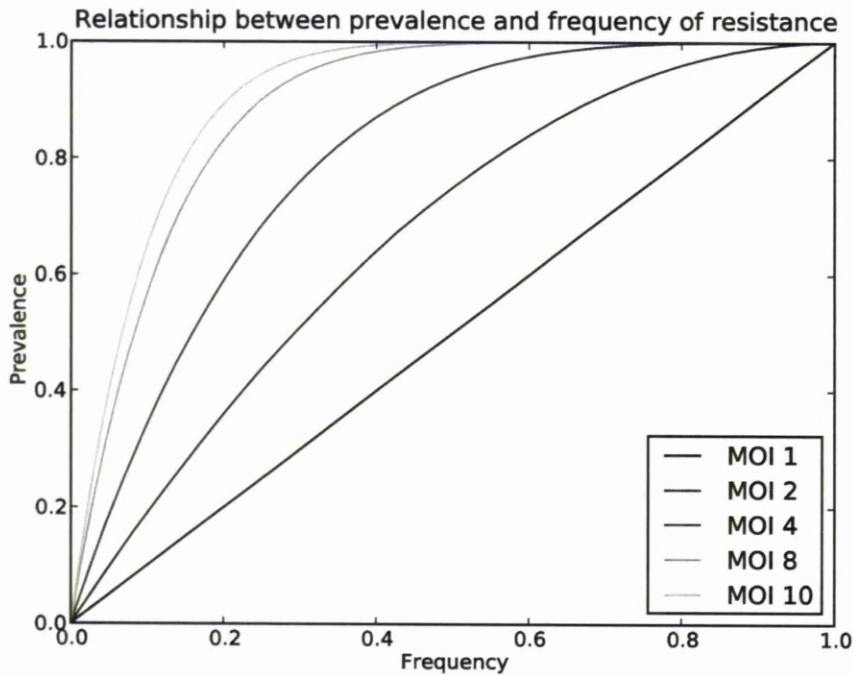


Figure 3.3: Relationship between prevalence and frequency of resistance with multiplicity of infection (MOI) in the single locus model. For the same frequency, higher MOI entails higher prevalence.

Model	MOI 1	MOI 2	MOI 4
Single locus	0	8	24
Full epistasis	0	5	23
Duplicate gene function	0	53	70
Full epistasis + DGF	0	12	31
SP-based	0	10	24
Two drugs	2	13	27

Table 3.2: Percentage of scenarios which stabilise at intermediate frequencies of resistance by MOI. MOI affects both effective fitness costs and the level of sexual recombination.

additional factor $s(i)$ could be added to this equation to incorporate other factors such as increased parasite clearance or lower gametocyte densities that may occur and be independent of MOI. When MOI is bigger than 1, it becomes an important factor in the loss of resistance when a drug is removed. Higher MOI increases the rate of spread of sensitive parasites and this effect is stronger at higher frequencies of resistance (data

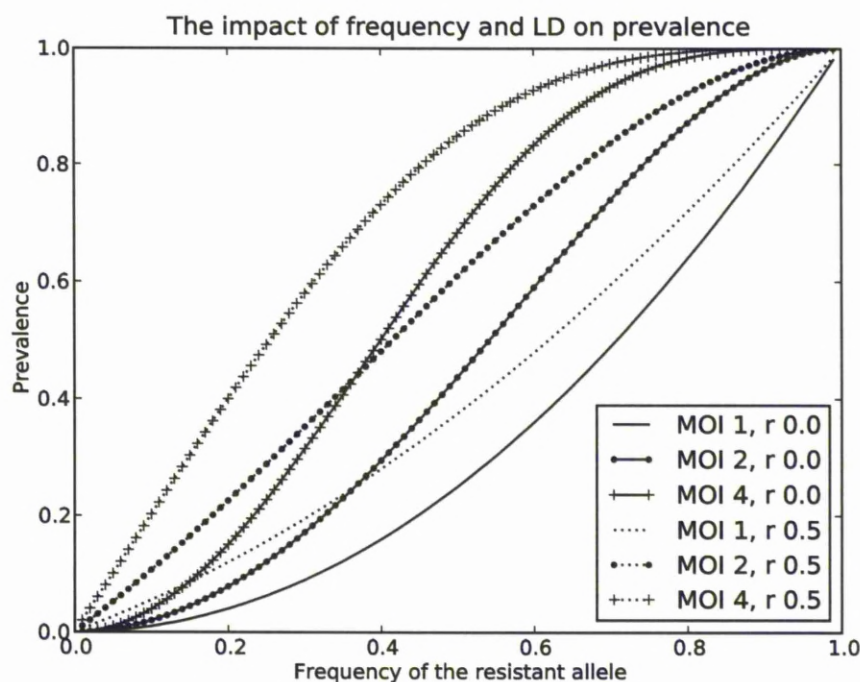


Figure 3.4: Prevalence as a function of the frequency of the resistant allele with full epistasis for linkage disequilibrium (r) of 0.0 and 0.5 for the full epistasis model.

not shown). This arises because both factors increase the probability that a sensitive clone will compete with resistant clones within untreated humans; the formers' superior competitive ability (they lack fitness penalties associated with resistant mutations) within humans helps drive their spread through the population.

3.3 Discussion

The methodology described above shows that it is conceptually straightforward to incorporate a flexible genetic basis of resistance into models of antimalarial drug resistance, and that this may have important qualitative implications for important factors such as the rate at which resistance evolves and whether it may stabilise at intermediate frequencies. This entails a slight redefinition of the concept of “resistance” away from hard-wired genetic determinism and towards a more subtle realisation that drug “resistance” is determined by both the parasite genome, and the “environment” of the infected host: two hosts might carry infections with exactly the same genotype but have different treatment outcomes depending on their immune status, pharmacogenetics and/or compliance to the recommended drug regimen. Chloroquine provides a useful illustration

of this. There was widespread “resistance” in Guinea Bissau, but when that country adopted a policy of doubling the CQ dosages, the problem of “resistance” largely disappeared (Ursing et al., 2007): the mutations, probably in *crt*, encoded resistance to normal levels of CQ but remained drug-sensitive when exposed to high levels. The presence of multiple epistasis environments in the host population (Table 3.1) reflects the importance of human immunity and pharmacological variables, such as drug quality and correct dosage. A key strength of this approach lies in its flexibility as it can also easily be used to incorporate other factors besides the main ones of immunity, pharmacogenetics and compliance explicitly discussed above. Other factors could be human genetic variation in malaria susceptibility and the role of residual drugs in driving resistance. Most antimalarial drugs have long half-lives and are frequently taken in many areas to presumptively treat any fever, with the consequence that a large proportion of people (up to 80%, e.g. Talisuna et al. (2002)) carry “residual” levels of drug from previous treatments. Most antimalarial drugs do not affect *P. falciparum* during its initial incubation in the liver and parasites emerging from the liver may encounter drug and be subject to “accidental” drug action which may be an important driver of resistance (Hastings et al., 2002; Hastings and Watkins, 2006). It is easy to envisage a plausible situation where full epistasis may describe survival to direct therapeutic treatment while in the bloodstream, asymmetric epistasis may describe resistance to high residual drug levels encountered on emergence from the liver and DGF describe survival to low levels of drug encountered after liver emergence. It would be straightforward to extend our approach to this situation by constructing a new scenario similar to those described above.

Fixation of sensitive alleles is far more likely to occur in the “two drugs” scenario (data not shown). This occurs because mutations encoding resistance to one drug offer no protection against the second drug but are deleterious in untreated individuals. If we extrapolate this result and assume that the sensitive alleles are more likely to be fixed if more drugs are used, and that the resistant mechanisms to all drugs involved are independent, then it should be possible to treat higher number of individuals before resistance spreads. This is consistent with the results presented elsewhere (Koella and Antia, 2003; Boni et al., 2008).

Simulations of drug removal following fixation of resistance allele illustrates the impact of untreated infections in the spread of drug resistance. Untreated infections delay the spread of drug resistance because sensitive infections are favoured in these hosts. This effect is more pronounced in higher transmission settings as MOI is high, increasing competition between clones, and is especially significant with higher frequencies of resistance. This highlights, and quantifies, a fundamental conundrum in drug deployment: an individual-centred, medical approach dictates that symptomatic patients be identified and cured, while drug policies aim to reduce the overall amount of drug used to minimise selection for resistance. These are not entirely incompatible considerations,

for example better diagnostics can reduce drug usage without reducing patient care, but it is an important component of models that they can quantify the impact of changes in drug deployment policies. Population genetics models also clearly highlight the tension between measures for control (defined as the reduction in incidence of the disease) and elimination (the reduction to zero of the incidence in a specific human population) of malaria. If control is the main objective, semi-immune, asymptomatic hosts provide a reservoir of sensitive parasites, where these are fitter than resistant parasites. From a control perspective having a reservoir of sensitive parasites is a positive development which slows the spread of resistance and increases the time a drug will remain effective. From an elimination perspective asymptomatic reservoirs have to be treated in order to completely remove the parasite from the host population and are a source of concern as they are difficult to detect in the human population. Any attempt at elimination, if unsuccessful, will probably have negative consequences for long-term control as selection against sensitive parasites in elimination policies will most probably increase the frequency of resistance against any drugs used in the elimination phase.

It is generally assumed that mutations encoding drug resistance pay a fitness penalty and are deleterious in the absence of the drug. Consequently, the frequency of resistant parasites should fall once that drug is removed from general use. This effect is important as reintroduction of the drug might be considered if the frequency of resistance to a certain drug drops to very low levels. For example CQ reintroduction as a partner drug in combination therapy, probably with artesunate, has been considered in Malawi (Laufer et al., 2006). Our results (data not shown) suggest the rate of fall will be faster in high transmission settings as a consequence of higher MOI and hence higher competition within hosts. This can be observed in the field where the return of Chloroquine sensitive parasites was observed after the introduction of the replacement drug SP, especially in high-transmission areas like Malawi (Kublin et al., 2003) and to some degree in Gabon (Schwenke et al., 2001) but less in low-transmission areas like Colombia and Venezuela (Cortese et al., 2002) suggesting that intensity of transmission might be a factor in decreasing levels of drug resistance. However these areas differ in many other aspects besides MOI (for example, migration from border countries or provinces with different drug deployment policies) and the results need to be interpreted with caution; for example, the fall in resistance was faster in areas of low MOI in Yunnan province, China (Yang et al., 2008), but the area shows highly heterogeneous patterns of transmission so other factors might have a substantial impact.

A large body of existing research predicts that once drug resistance arises it will spread rapidly to fixation. Hastings (2006) postulated that forces driving resistance (genetic recombination, intrahost dynamics, natural selection) vary with frequency and could cancel out to allow stable intermediate frequencies; he used a model of full epistasis and showed that stable intermediate frequencies could occur, but was fairly uncommon, especially for lower MOI. In contrast, many field studies suggest stabilisation is relatively

common; for example, for SP in Malawi (Plowe et al., 2004) and Tanzania (Pearce et al., 2003) and for CQ in Eastern Sudan (Babiker et al., 2005) and Guinea-Bissau (Ursing et al., 2007). The more biologically realistic “Full Epistasis + DGF” and “SP-based” models of host heterogeneity described here exhibit an increase in intermediate stable frequencies especially with lower MOI (Table 3.2). In summary, realistic modes of gene interaction are more compatible with field observations, demonstrating the importance of accurately modelling gene interactions and also suggesting the need of further empirical research on the basis of drug resistance.

Previous models based on an assumption of full epistasis, always predict strong positive linkage disequilibrium (Dye and Williams, 1997) here we show that LD depends on the mode of gene action (Figure 3.2), which is not well understood for most drugs and which will, as stressed above, depend on the selection environment of the treated human. This has implications for interpretation of field data. Studies have investigated LD between loci in the belief that significant positive LD would be indicative that both loci are important for encoding resistance, the most obvious examples being *mdr* and *crt* loci in CQ resistance. This is true if full epistasis is required, but Figure 3.2 shows that “negative” results (i.e. absence of LD or negative LD) cannot be taken as evidence that they do not have a joint role in determining resistance. LD also has a strong impact on genotype inference and on the ability to accurately calculate the frequency of resistance from prevalence (Figure 3.4). Many clinical studies present the prevalence or frequencies of each locus separately, and normally no attempt is made to report multilocus genotypes because it is often impossible to determine a multilocus genotype (i.e. linkage phase in the population genetic terminology) when $MOI > 1$. The frequency of multilocus resistance genotypes, and the direction and extent of LD between the loci, provides clues as to the underlying genetic mechanisms of resistance so a strong case could be made for future studies to attempt genotype inference and LD estimation from multiple infections (Hastings and Smith, 2008).

Intensity of transmission determines MOI, which determines the proportion of individuals carrying resistant infections (i.e., the prevalence of resistance). This has two main implications for comparative analysis of clinical studies from areas with different transmission intensity, or for temporal studies from the same area if the transmission rate has changed:

1. The potential confusion between frequency and prevalence of resistance should be avoided. Different transmission settings change the relationship between frequency and prevalence and higher MOI clearly entails higher prevalence for the same frequency of resistance.
2. For longitudinal studies where transmission has decreased, an observed drop in prevalence of resistance does not always reflect decreased frequency of resistance. Any conclusion that cutting transmission decreases the frequency of resistance

should be carefully evaluated as such observation could be explained, partially if not totally, by a lower MOI generating a smaller prevalence for the same frequency (Figure 3.3). The relationship becomes much more complex when more than one locus is involved as LD and epistasis also affect prevalence (Figure 3.4).

Our theoretical conclusions on the relationship between transmission, frequency and prevalence are consistent with a recent study (Hastings et al., 2010) which analysed field data from Tanzania and Papua New Guinea.

P. falciparum biology has several features that preclude simple analysis using standard population genetics equations. Studying the implications of parasite inbreeding, epistasis between drug resistance loci, and heterogeneity in human host “environments” will allow a better understanding of the dynamics of resistance spread of *P. falciparum* malaria. This explicit and flexible framework of gene action can be used in the future to study deployment strategies for the new generation of antimalarial drugs, the artemisinin combination therapies (ACTs), such as simultaneous drug deployment policy proposed by Boni et al. (2008) and to understand the epidemiological consequences of different genetic mechanisms of resistance. A notable omission from previous analyses has been the recognition that resistance depends on both parasite genotype and the human “environment” in which treatment occurs. Incorporating this effect should allow enhanced understanding of parasite population genetics of drug-resistance genotypes allowing us to identify public health deployment practices that may minimise selection for resistance and ultimately to better mitigate the public health impact of malaria.

Four

The promise and dangers of recent antimalarial deployment policies

Tiago Antao

Abstract

Malaria is one of deadliest infectious diseases, imposing a significant health, social and economical burden, particularly on poverty-stricken countries and populations. Antimalarial drug resistance is pervasive and several strategies have been proposed to reduce the spread of drug resistance and extend the useful therapeutic life of existing therapies. Recent research has suggested that using multiple first-line therapies (MFT), instead of a single-drug policy, might delay the emergence and spread of drug resistance. Here we compare different drug deployment policies using several realistic population genetics models of the spread of drug resistant malaria focusing on the spread of multidrug resistance. We simulate realistic models of genetic interaction, immunity, treatment compliance and the impact of transmission intensity on inbreeding. Our results suggest that MFT policies can delay the spread of drug resistance, including multiple resistant genotypes, if resistance levels are maintained inside World Health Organisation recommendations. This work stresses the importance of compliance with treatment guidelines as incomplete treatment can potentially increase the spread of resistant infections. We also show that association between loci, measured by linkage disequilibrium, is fundamental to understand the dynamics of spread of resistance as it is the main driver for the spread of multidrug resistance.

Malaria is a major public health concern, as one third of the human population is estimated to be exposed to the threat of the most virulent form, *Plasmodium falciparum* (Snow et al., 2005). Antimalarial drug resistance has emerged as one of the major challenges facing malaria control as it became widespread to most first line therapies based on monotherapies (Olliaro, 2005). Treatment failure is now being observed even for Artemisinin based combination therapies (ACTs) (Rogers et al., 2009). It is now mandatory that antimalarial drugs be deployed as combinations because modeling

suggested this was a much more sustainable policy than using monotherapies (World Health Organization, 2006). Historically, in the era of monotherapies there was little room for policy manoeuvre: one drug, Chloroquine (CQ), was used before it became ineffective when Sulphadoxine-Pyremethamine (SP) was deployed until it too became ineffective. We are presently in a slightly more advantageous (but still tenuous) position compared to the monotherapy era because we have several effective combination therapies available based around Artemisinins using Amodiaquine, Lumefantrine, Mefloquine, Piperaquine and SP (where the latter is still effective) and even CQ as a possible partner-drug (Laufer et al., 2006). It therefore seems sensible to consider how best to deploy these ACTs for maximal therapeutic lifespan. Current drug deployment policies, hereafter termed “sequential”, make use of a single first line therapy which is used until the level of treatment failure rises above an acceptable level, at which time the drug is replaced with a new treatment. A different strategy would be to deploy multiple first line therapies (MFT) simultaneously, resulting in humans and thus infections being treated with different drugs.

The benefits of MFT are intuitively obvious. It is generally accepted that it is more difficult for organisms to evolve in a heterogeneous environment, in this case caused by different drug treatment, than in a homogeneous treatment environment associated with single drug use. The benefits become even more pronounced if fitness penalties are associated with mutations encoding drug resistance. Koella and Antia (2003) recognised a “tipping point” in this situation whereby the advantage of a mutation in encoding drug resistance in treated humans is balanced against its fitness cost in untreated individuals; resistance will only spread if its benefit outweighs its cost. MFT reduces the proportion of infections treated by each drug, so, in principle, it could reduce a mutation’s advantage below this tipping point and the mutations would be selected out of the parasites population. Under a best-case scenario this could prevent resistance spreading to any of the drugs and could even reverse its spread once started (Hastings and Donnelly, 2005). Even if the cost/benefit does not cross the tipping point, the use of MFT will still be advantageous because reducing drug use has a disproportionate impact on selection for resistance. For example halving drug use may reduce selective advantage 2.5 fold (Babiker et al., 2009).

However, there is also a potent intuitive threat posed by MFT: the parasite has an obligatory sexual phase inside the mosquito so infections that are, individually, resistant to a single drug can recombine to create multi-drug resistance. A worst-case scenario is that resistance mutations become genetically associated so that multi-drug resistant parasites rapidly spread that are resistant to all drugs in the MFT arsenal and completely undermine the benefits of MFT. If true this would suggest that sequential use of drugs might be a better policy. This effect, termed linkage disequilibrium (LD), is known to be an important dynamics in the evolution of drug resistance (Dye and Williams, 1997; Antao, 2010). Note that LD does not require resistance to be physically linked on the same

chromosome; it is a statistical association caused by resistance mutations “hitch-hiking” with each other because of their mutual benefit in withstanding all drug treatments. It would be impossible to recommend the use of MFT without explicitly considering the risk posed by multi-drug resistant infections (see current concerns about, for example, multidrug resistant Tuberculosis (Zignol et al., 2006) and Methicillin-resistant *Staphylococcus aureus* (MRSA) (Enright et al., 2002)).

A second potential threat to MFT arises because most of the current batch of anti-malarial drugs (except the Artemisinins) have long half-lives and may persist at active concentrations in humans for weeks after treatment. This effect is known to drive resistance (Hastings and Watkins, 2006) because parasites emerging from the liver may encounter, and have to survive, residual drug levels persisting from previous treatments. Many patients in areas of high drug use may have residual levels of drug so MFT will often morph into a type of *de facto* combination therapy whereby parasites must be simultaneously resistant to Drug A to survive residual drug levels and establish an infection, and also resistant to Drug B to survive later therapy by that drug. It is therefore important to understand the interactions of drug deployment policies and residual drug levels, especially on the spread of multi-drug resistance and, more optimistically, whether MFT may be an inexpensive method of harnessing the additional known benefits of combination therapy.

A previous evolutionary-epidemiological study (Boni et al., 2008) suggests that MFT would result in longer overall periods of drug effectiveness, but only briefly considered the threat posed by the emergence of multi-resistant genotypes through the *P. falciparum* sexual reproduction phase; in particular they assumed random breeding among *P. falciparum* parasites within the whole parasite population whereas we explicitly allow recombination only between *P. falciparum* clones co-infecting the same human. This is important because, by definition, only resistant genotypes survive treatment so recombinational loss of resistance will be reduced or eliminated in infections transmitted from treated humans; this will increase the genetic stability of multidrug resistance greatly enhancing its spread.

Any endorsement of the use of MFT should consider the risk posed by creating LD and multi-drug resistance. We also take the opportunity to include several other factors omitted from the previous MFT study in particular: the impact of different modes of genetic interactions (epistasis) between loci conferring resistance; the impact of intra-host dynamics and competitive release (Hastings, 2006; Wargo et al., 2007). In this study we construct explicit population-genetic models to investigate how parasite population structure and genetic mechanisms of resistance influence the outcome of different drug deployment policies. Key questions to be addressed are: How do different types of selection pressure influence the spread of resistant genotypes? What is the impact of control and elimination programmes which decrease transmission? If MFT is to be tried, should we start in high or low transmission areas? Is MFT always the best strategy in

all epidemiological settings? Do any of these strategies increase the spread of multiple resistant clones? As MFT might already be a *de facto* policy due to the widespread availability of multiple drugs through informal sector (Bate et al., 2008), should this be encouraged or suppressed? The model presented is general enough to study many different parameters but we will concentrate on scenarios of most practical importance and realism as regarding *P. falciparum* population biology.

4.1 Modeling and theory

We used ogaraK (Antao and Hastings, 2011a), a population genetics simulator of the emergence and spread of drug resistant *P. falciparum* that incorporates multiple loci, sexual recombination, LD, differing levels of multiplicity of infection (MOI, see below) and different genetic modes of resistance. Here we present a summary of the relevant features of the application and the parameters used in this study.

Population genetic models for *P. falciparum* have to address several non-standard features of its biology. *Plasmodium* malaria parasites are haploid and reproduce asexually in humans and are briefly diploid in the mosquito vector, where they reproduce sexually. The number of simultaneous infections in a human is termed the Multiplicity of Infection (MOI) which typically ranges from 1 to 12. The number of mating options inside the mosquito are dependent on the MOI of the human providing the blood meal. Different infections might have different resistance profiles. Sexual reproduction in the mosquito entails recombination but self-fertilisation (i.e., selfing) occurs frequently as mosquitoes might ingest parasites having a single clone (Arnot, 1998) (mosquitoes feed approximately every three days so mating between parasites in blood meals obtained in separate bites is assumed to be impossible). This creates an environment which departs from standard population genetics' models, namely that mating can only occur between the small number of different parasite clones within a blood meal. A crucial difference between areas of high and low malaria transmission is the MOI: repeated sequential infection in areas of high transmission intensity leads to the average number of different clones being higher, so mosquitoes frequently ingest unrelated parasites leading to lower levels of selfing and inbreeding (Anderson et al., 2000a). MOI is thus a proxy for transmission intensity because higher transmission increases MOI due to repeated sequential infection.

We track the spread of resistance over a time-scale of 200 parasite generations. Each generation encompasses a *P. falciparum* life cycle – mosquito-human-mosquito – which is likely to be around 5 per year so the simulation lasts 40 years which should encompass all the timescales likely to be considered in long-term planning. We simulate the genotype of each infection assuming that all loci are physically unlinked, a realistic assumption for loci known to encode *P. falciparum* drug resistance (Osman et al., 2007). We also make the simplifying assumption that there is no cross resistance among drugs

(i.e. a single mutation cannot encode resistance to more than one drug), but note that this is not applicable to all known cases (Price et al., 2004). Each loci can have two alleles: resistant and sensitive. Resistant alleles will incur a fitness penalty if they are not required for survival (i.e. in untreated hosts, in hosts treated with a drug for which the mutation cannot encode resistance or in cases where the epistasis mode does not require all mutations). Genotypes with multiple mutations incur a multiplicative fitness penalty. More than one locus can be involved in resistance to a single drug. Importantly, where more than one locus encodes resistance to a single drug, we investigate different epistasis modes among loci (Antao, 2010): Duplicate Gene Function (DGF) where a resistant allele at any locus is sufficient to confer resistance to a drug; Full Epistasis (FE) where resistant alleles have to be present at all loci to confer resistance; Asymmetrical Epistasis where one locus is more important (for example SP resistance where mutations in *dhfr* have a more substantial role than *dhps* (Olliaro, 2001)). Full epistasis is deemed a “strong” mode as it requires all mutations for resistance, conversely DGF and Asymmetrical Epistasis are called “weak”. Parasite multi-locus genotypes can vary from sensitive to all drugs to resistant to all drug treatments available and having all mutations (multi-resistant).

Infections are assumed to have equal infectivity and mate at random inside the mosquito so that if there are n clones in a human then selfing rate is $\frac{1}{n}$ and outcrossing rate $1 - \frac{1}{n}$, and competitive release is assumed to occur. Parasites in infected humans can be untreated (an environment where sensitive infections are fitter), treated in humans who have little or no host immunity (where a weaker epistasis mode suffices to confer resistance) and treated in humans who are semi-immune to infection (requiring the strong mode of resistance).

We assume that resistance to all therapies exist at low frequencies (0.1%) at the onset of the simulation. The emergence of resistance and its implications on therapy effectiveness have been studied elsewhere (Boni et al., 2008; Pongtavornpinyo et al., 2009). Our approach is complementary as we try to understand the spread, rather than the origin, of existing resistance. Furthermore resistance to most drugs is now widespread, and has probably emerged also for artesunates (Dondorp et al., 2009) and in many cases *de novo* resistance has arrived to a human population via migration (Roper et al., 2004), not local mutation.

We considered three drug policies: (i) sequential application, modeling the common antimalarial deployment strategy where a therapy is replaced by another when treatment failure becomes too high; (ii) multiple-first line therapies (MFT) where several drugs are made available allowing patients and clinicians to randomly choose which one to use and (iii) combination therapy where all available drugs are given simultaneously to each patient. The combination therapy model is inspired in Tuberculosis policies where all available drugs are used in a single patient (Crofton et al., 1997), and should not be

confused with Artesunate Combination Therapies (ACTs) where a Artesunate derivative is supplied with a single partner drug.

In order to compare policy duration we assume that a sequential policy lasts until the last drug is removed from circulation. A drug is replaced as soon as an average of 10% treatment failure is observed. For MFT and combination therapy we assumed that a policy stops being effective as soon as 10% treatment failure is observed. This figure of 10% was chosen because the World Health Organization (WHO) recommends a change of treatment regimen when cure rate falls below 90% (World Health Organization, 2006).

We simulated sequential application, MFT and combination assuming two drugs are available and that resistance is encoded in two different loci per drug (i.e a genotype consists of four independent loci) using the following set of scenarios:

Full epistasis: Parasites must possess resistant alleles at both loci to survive treatment with that drug (for example parasites must have resistance mutations present at both *dhfr* and *dhps* to survive SP treatment).

DGF: Parasites with resistant alleles at either locus will survive treatment with that drug (for example parasites with resistance mutations at either *dhfr* or *dhps* will survive SP treatment).

Mixed mode: Half of the treated humans have no host immunity and/or take low drug doses. Parasites in these humans will be able to survive treatment if they have a resistant allele at the most important locus (asymmetrical epistasis); this scenario reflects observations in both SP (where *dhfr* is more important than *dhps*) and CQ (where *crt* is more important than *mdr1*) (Antao, 2010). The other half of the human population is assumed to be semi-immune and/or take high drug doses, therefore only parasites that have resistant alleles at both loci will survive treatment (Full epistasis).

For all simulation scenarios we varied the fitness penalty per resistant mutation and the amount of drug usage (defined as the percentage of infected humans that were treated) both were between 0 and 100% in increments of 2%. Only results with fitness penalties below 20% are reported here as large values are not realistic. We simulated 4 different MOIs: two simple models with MOI fixed at 2 (low transmission) and 4 (high transmission) and two more realistic scenarios, one modeling low transmission where 50% of human hosts had a single infection and the other half had 2 infections and another modeling high transmission where MOI followed a Poisson distribution with a conditional mean of 2.3 truncated at a maximum MOI of 7 (Hastings, 2006). The simpler MOI models qualitatively capture the results of the more complex ones (Antao and Hastings (2011b), supplemental information), so we opt to present only the results pertaining to the simple MOI distributions. We repeated the above scenarios with 3

drugs but the results were qualitatively similar to the results with 2 drugs; we therefore only present results for 2-drug simulations except when comparing the importance of the number of drugs used.

4.2 Results

Table 4.1 summarises the impact of key factors on the spread of drug resistance and table 4.2 the fundamental consequences of different drug deployment policies. Below we compare the three deployment policies and detail the results for each key factor.

Factor	Impact
Drug use	↑ resistance spread
Fitness cost	↓ resistance spread
MOI	↑ competition → ↓ resistance spread ↑ recombination → ↓ LD ↑ asymptomatic infection → ↓ drug use
LD (r)	↑ resistance spread (except if DGF) → ↓ resistance spread
Genetic mode	↓ resistance spread (full > mixed > DGF > none)
Compliance	↑ strength of genetic mode
Immunity	↑ strength of genetic mode

Table 4.1: The impact of key factors in the spread of drug resistance. The left side of the table lists important factors involved in the spread of drug resistance. The right side describes the impact of increasing that factor. The table not only summarises the impact of several factors but also demonstrates the complex web of relationships between factors and impacts. This table is applicable for realistic frequencies of resistance according to World Health Organisation policy, i.e. where resistance is considerably below 50%.

Policy	Impact
Combination therapy	Longest useful therapeutic life Extreme increase in multi-drug resistance
MFT	Long useful therapeutic life Very low multi-drug resistance
Sequential application	Medium useful therapeutic life Low multi-drug resistance

Table 4.2: Summary of the impact of MFT, sequential application and combination therapy on useful therapeutic life and the spread of multi-resistant genotypes assuming drug usage typical in control scenarios.

Policy comparison

MFT policies last longer than equivalent sequential policies for low to medium drug usage. For higher drug usage, sequential application performs better, but in this case the difference between policies is minor (Figure 4.1). The dynamics of genotypes that encode resistance to drugs not in use in sequential application is fundamental: when a new drug B is introduced into the sequence, it will compete against a mixture of parasites that are either sensitive or resistant to drug A, these resistant forms will be less competitive as they have mutations that only decrease fitness, thus resistance to drug B will spread faster (Boni et al., 2008). However, a compensatory effect will occur with higher drug usage: parasites that are spontaneously resistant to drug B will be eliminated at a higher rate when drug A is in use, therefore when the drug B is introduced the basal frequency of infections resistant to that drug is highly reduced. The relative benefit of MFT against sequential usage is then a balance between these two processes.

Combination therapy shows a similar profile to MFT, while lasting slightly longer. The fundamental qualitative difference between combination therapy and the other policies lies on the spread of multi-resistant infections: the frequency of the multi-resistant genotype will be small at the end of the useful therapeutic life with MFT and sequential application, whereas with combination therapy the multi-resistant genotype will usually be above 50%. Figure 4.2 shows the fraction of resistant genotypes which are multi-resistant at the end of policy duration for sequential application, MFT and combination therapy. These results are intuitively expected as in both MFT and sequential application, the multi-resistant genotype is never the most fit in all human hosts because in any human this genotype pays fitness costs for carrying resistance mutations that are not required for survival in the host; consequently multidrug resistance will only become frequent, through random association, when resistance to all drugs is very high. On the other hand, with combination therapy the multi-resistant genotype (or genotypes, in the case of weaker epistasis modes) is the only one that can resist treatment. As multiple resistance is a serious problem with many infectious diseases and the system behaviour has been shown to be function dependent (i.e., the behaviour is qualitatively different with low and high resistance frequencies (Antao, 2010)), we also made the same comparison for a much higher threshold of resistance (50%) before a drug is removed. Simulations show that MFT multi-resistant pattern shifts to an intermediate between sequential application and combination therapy (Figure 4.3).

Multiplicity of infection and Linkage disequilibrium

Recombination will reduce any statistical association between resistant alleles at different loci and the frequency of resistance qualitatively influences the impact of recombination. One of the fundamental assumptions in any malaria model of resistance is that the

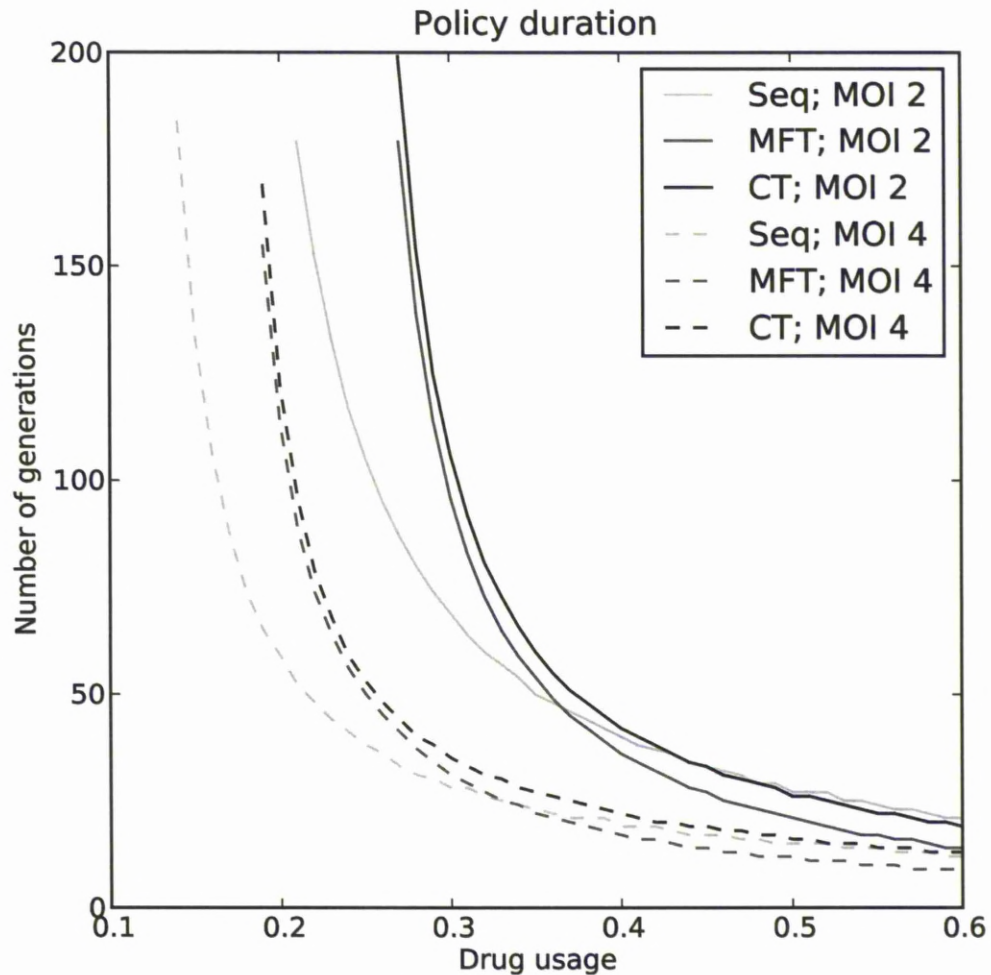


Figure 4.1: Impact of policy and MOI on useful therapeutic life; the latter is plotted on the Y axis and is defined as the number of parasite generations that elapse before overall drug failure rates reach 10%. It is plotted as a function of drug usage (the proportion of infections that are treated) on the X-axis. The chart compares combination therapy (CT), MFT and sequential (Seq) deployment policies with MOIs of 2 and 4 and with a fitness penalty of 10%. MFT and CT perform better than sequential policies at lower drug usage and marginally worse with high drug usage. All three deployment policies last longer with lower MOIs.

frequency of sensitive alleles is greater than 50% (as WHO policies postulate efficacy levels above 90% (World Health Organization, 2006)). If a clone is resistant to one drug, a recombination event involving a different clone will probably generate offspring that are only resistant to the same drug, as the other clone is probably sensitive (due to the assumption of low frequency of resistance). Linkage disequilibrium patterns can change with different epistasis modes and drug policies. Figure 4.4 shows the LD (r)

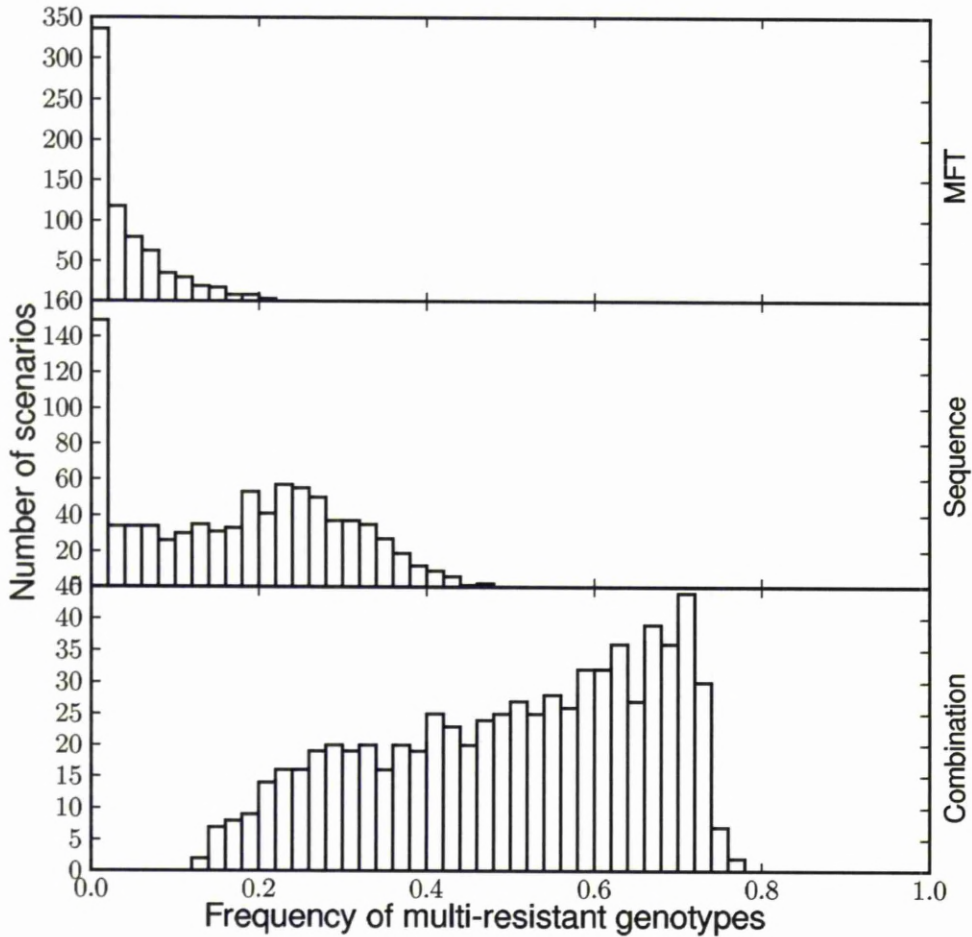


Figure 4.2: The distributions of multi-resistant genotypes at the end-of-life for MFT, sequential application and combination therapies. MFT and sequential application show similar patterns i.e. most genotypes are not multi-resistant. Combination therapy is qualitatively different as most genotypes at policy end of life are multi-resistant. The distribution includes all simulations with a fitness penalty below 20% and drug usage below 60%.

for both policies assuming full epistasis or DGF the whole 200 generations simulated. Both the signal and magnitude of LD varies with epistasis and policy as it is positive in full epistasis for loci involved in resistance to the same drug and negative for DGF. From an empirical perspective, all signals and intensities of LD are plausible depending on policy and epistasis mode (Antao, 2010). Note that positive LD indicates resistance alleles are found together in the same parasite genotypes more often than expected by chance, and negative LD indicates that they are associated less often than expected.

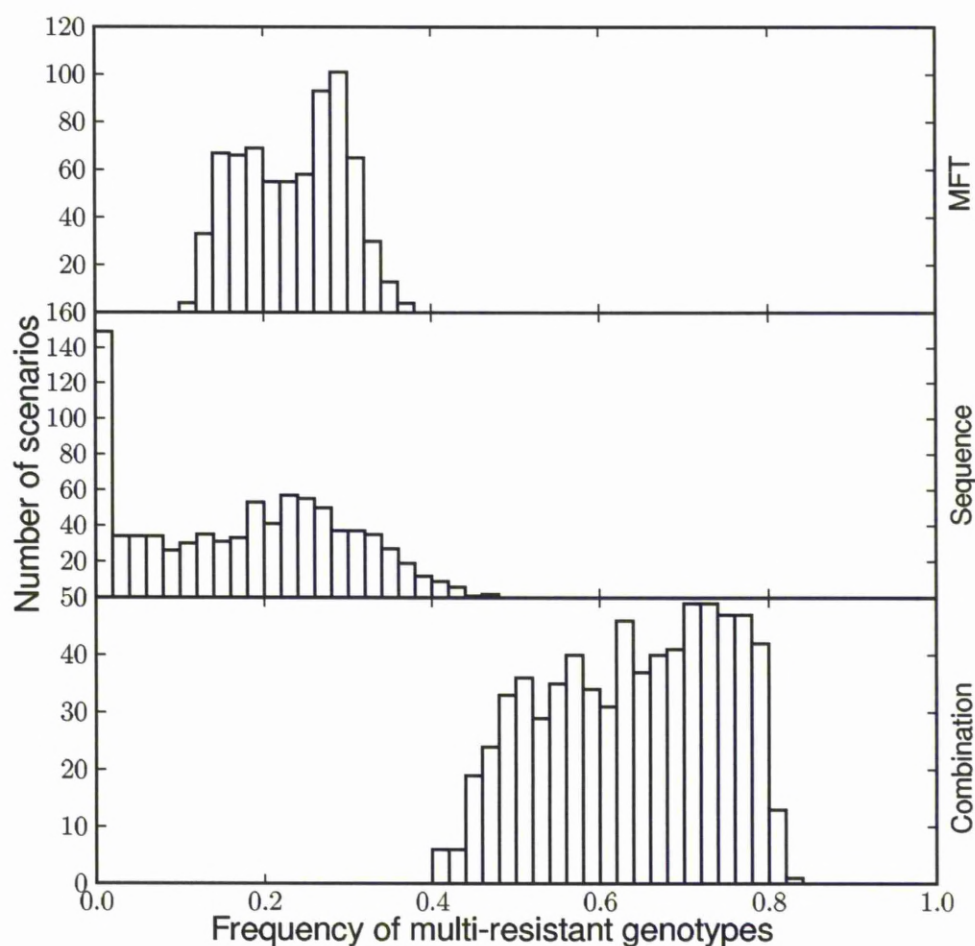


Figure 4.3: The distributions of multi-resistant genotypes at the end-of-life for MFT, sequential application and combination therapies. End of life is assumed to occur at 50% prevalence (compared to 10% in figure 4.2) but all other parameters, abbreviations and simulation procedures are identical between the two figures.

MOI is an important factor in determining the length of effective drug use as all policies last longer with lower MOI (assuming resistance is present at the onset of simulations). For the same frequency of resistance, higher MOI increases the proportion of treated humans with one or more resistant clones (the “prevalence” of resistance) (Antao, 2010) allowing a higher frequency of resistant genotypes to be transmitted to the next generation. However, this effect is countered, in higher MOI settings, by breakage of LD that will occur in untreated individuals by outcrossing between resistant and sensitive infections; and outcrossing increases with higher MOI. The lesser importance of breaking LD compared to prevalence in establishing the spread of resistance can be

intuitively understood if we assume that the fitness penalty for carrying mutations is zero. If there is no fitness penalty, competition in untreated individuals is absent and in this case untreated environments will tend towards linkage equilibrium. Thus, the breaking of association between resistance alleles will only happen in cases where there is strong linkage between them. In fact, if linkage was negative (as in DGF, Figure 4.4), untreated individuals would contribute to increase the association between resistance alleles.

Epistasis

The spread of resistance is faster in environments with weak epistasis modes. Figure 4.5 exemplifies this as both MFT and sequential application last longer with full epistasis than with mixed mode. In fact, the modes of resistance are more important in determining policy duration than the policies themselves. In mixed mode, the main locus is enough to confer resistance in half of the environments treated with a drug so there is no need for association with a second locus. This is quantitatively more important for high fitness penalties: The second locus encoding resistance to a drug is not needed in half of the drug treatments (in contrast to full epistasis where it is always needed), therefore it is often deleterious even in the presence of a drug.

Number of drugs used

The number of drugs used does not qualitatively change, from this policy comparison (Figures 4.6 and 4.7): using more drugs will increase the useful therapeutic life proportionally. The “tipping point” where the advantage of mutations encoding drug resistance increases, allowing increased drug usage without significant resistance spread.

4.3 Discussion

Our results mostly confirm, from a perspective based on parasite population genetics, previous results on the possible effectiveness of MFT policies (Boni et al., 2008). MFT lasts longer than sequential application with low to medium drug usage. As semi-immune individuals are much more common in high MOI/transmission scenarios (as repeated infection leads to the development of immunity (Langhorne et al., 2008; Rogerson et al., 2010)) the expected fraction of untreated individuals will be higher, i.e., drug usage will be lower. This observation suggests that MFT might be more efficient in high transmission settings because lower drug usage benefits MFT policies.

MFT and sequential application show similar patterns of spread for multi-drug resistant genotypes up to levels of resistance assumed in WHO policies. In both policies, end-of-life analysis shows the relative low frequency of multi-drug resistant genotypes, with MFT showing slightly more favourable patterns. This is in contrast to the end-of-

life profile of combination therapies. However, if the frequency of resistance increases substantially above WHO standards than the MFT of multi-resistance profile shifts considerably (Figure 4.3). Therefore a realistic assessment of the ability to maintain resistance within WHO policy limits should be a fundamental decision guideline regarding the introduction of MFT.

Our sequential application analysis did not allow the option of re-using a drug after all possible therapies have been exhausted. Drug rotation, at the end-of-life of a sequential strategy, may reduce the frequency of resistance genotypes of the first drug used compared to the moment where the drug was removed. This result is compatible with the proposal for reintroduction of CQ as a partner drug for Artesunate (Laufer et al., 2006): our predictions are consistent with empirical results seen in Malawi where CQ resistance is severely depressed after the replacement with SP. This observation is highly dependent on the fitness penalty accrued by resistance loci, and these can vary from drug to drug. For instance, the loci involved in CQ resistance seem to incur higher fitness penalty, than SP resistance loci (Babiker et al., 2009). Therefore the option of re-introducing a drug (especially as a partner to an Artesunate based therapy) is highly dependent on a per case fitness penalty.

Residual drug levels have little impact on sequential policies because the residual and treatment drugs are identical. They do have an effect in MFT policies as parasites may have to survive one drug at residual levels, and survive treatment by another drug which, as noted in the Introduction, constitutes a type of combination therapy. We have not directly modeled this effect for several reasons. The proportion of people with residual drug treatment depends on the overall drug use which consists of treatment against malaria infection, and of presumptive treatment of people who have malaria symptoms (e.g. fever) but are not actually infected. Presumptive treatment is much more common in high transmission (high MOI) areas with poor clinical diagnosis so higher MOI settings show a positive correlation, and hence confounding, between lower therapeutic drug usage (due to immunity) but higher residual drug levels (due to high presumptive treatment). Residual drug levels may require weaker epistatic models than therapeutic drug use. Recent attempts to improve diagnosis using rapid diagnostic tests may greatly reduce presumptive drug use. Finally, different MFT implementations can alter the effects of residual drugs: if an MFT implementation can assure that the same individual is treated with the same drug over-time (either by tracking each individual treatment history or, more pragmatically, by having different therapies at different close geographical locations) then the impact of residual drug levels will be much reduced. Consequently we prefer an indirect argument of the effects of residual drugs on MFT: the difference between MFT and CT was small and CT was generally slightly better at maximising therapeutic lifespan (e.g. Figure 4.1) so we conclude, while noting the threat of multidrug resistance (Figure 4.2), that the fact that MFT may operationally merge into a type of CT would not undermine its deployment and may even act as a

kind of cost effective way of harnessing the benefits of CT without the financial and technical penalties of having to co-formulate drugs into a CT.

As with the fitness penalty (and its consequences on drug reintroduction), several assumptions should be scrutinised on a case by case basis. For instance most analyses and discussion was made assuming a range of fitness penalties per loci. This was carefully scrutinised through comparisons with simulations with no fitness penalty. Other assumptions do require future study: (i) if several different ACTs are used then there is clearly a partially shared resistance basis (on the Artesunate derivative) and further work is needed in modeling drugs with partially shared resistance; (ii) some drugs, notably Chloroquine and Amodiaquine (Sa and Twu, 2010), do share the same resistance loci but the mechanism of resistance seems to be different – even opposite and (iii) residual drug levels which play a critical role in the emergence and spread of resistance (Hastings et al., 2002). It is important that policy makers understand the limitations and assumptions of this (and any other) models of resistance. It is also important to note that current control and elimination agendas are explicitly aimed at identifying and treating all malaria infections (i.e. severely increasing drug usage) (Antao, 2011) and that our results regarding useful therapeutic life are partially reversed with high drug usage. Nonetheless, the difference at higher drug usage is relatively smaller than the difference at low to medium drug usage.

From a practical perspective a distinction should also be made between formal policy (which currently is sequential application virtually everywhere) and the pragmatic realities of different countries and regions. For instance, in several scenarios, while ACTs are the *de jure* policy, the private and informal sectors still distribute CQ and SP (Bate et al., 2008), therefore the *de facto* field reality is indeed MFT.

Compliance with treatment guidelines is fundamental to delay the spread of drug resistance. We presented different epistasis modes as modeling varying immunity profiles, but epistasis can also reflect treatment compliance (Antao, 2010). Full compliance is modeled by strong epistasis (full treatment forces the parasite to have all resistant alleles in order to survive) and poor compliance is modeled by weaker epistasis modes. Figure 4.5 clearly shows that stronger epistasis (full compliance) allows both policies to last longer. Our model provides strong support for the importance of full compliance and proper dosaging as spread is clearly slowed.

4.4 Conclusion

Regarding a comparison of policies, our results are mostly consistent with previous research. Most notably we confirm MFT to out-perform the standard policy of sequential application for realistic model parameters. Our results suggest that the impact of MFT on the spread of multiple resistant genotypes is negligible or even slightly better than sequential application as long as resistance is in within WHO guidelines. Importantly,

we do conclude that widespread availability of multiple ACTs through the informal sector is not an immediate cause for alarm (although Artemisinin monotherapies or low quality drugs should definitely be suppressed) and that countries may reasonably choose to deliberately employ a policy of MFT within the public health sector.

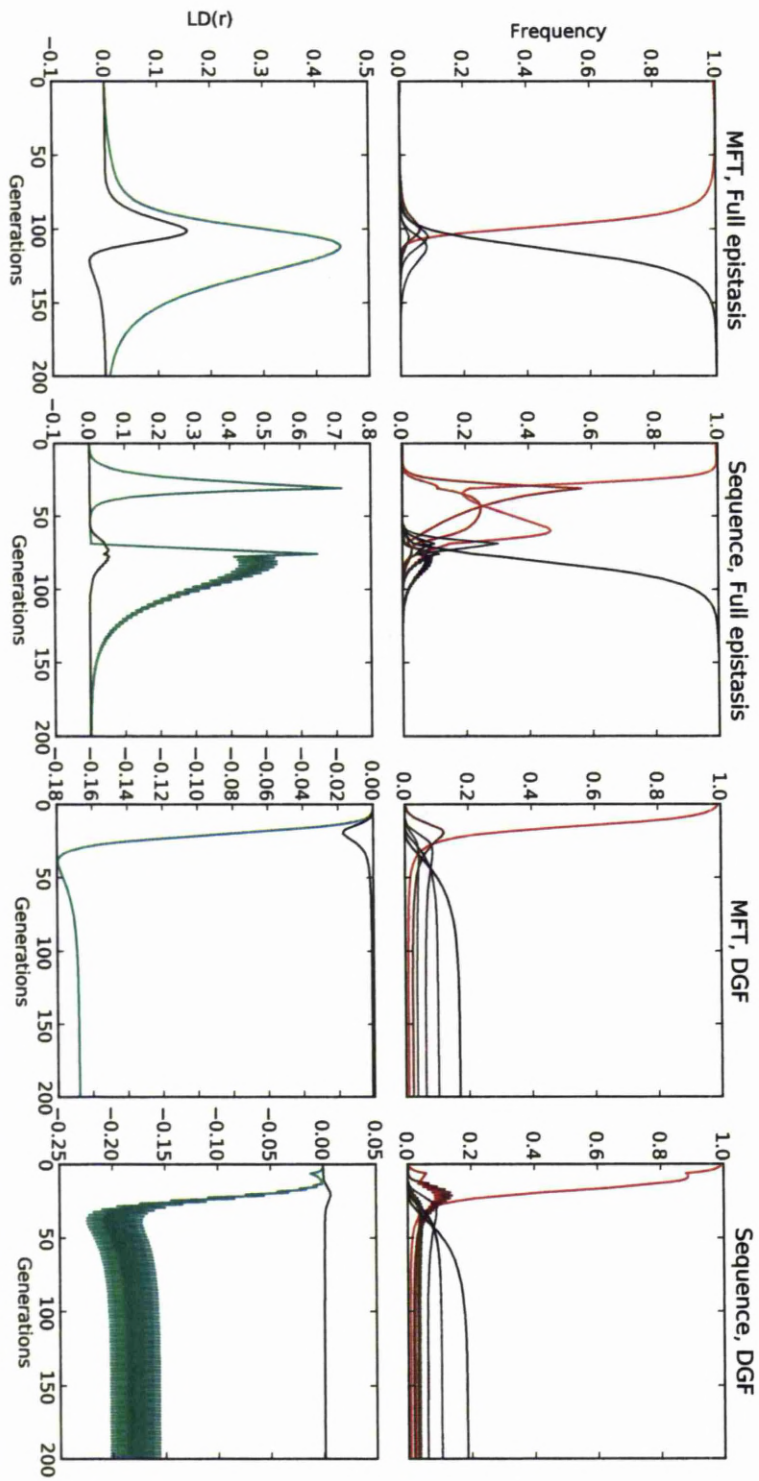


Figure 4.4: Frequency and linkage disequilibrium (r) with varying policies and epistasis modes for a drug usage of 30% and fitness penalty of 10%. Each column depicts a different policy (MFT or sequential application) and epistasis mode (full epistasis or DGF). The top row shows the frequency over time of all genotypes (a darker tone implies having more resistant alleles). The bottom row plots LD between two loci encoding for the same drug (green) and two loci encoding resistance to different drugs (black). Note that the Y scale for LD (r) is different on each column.

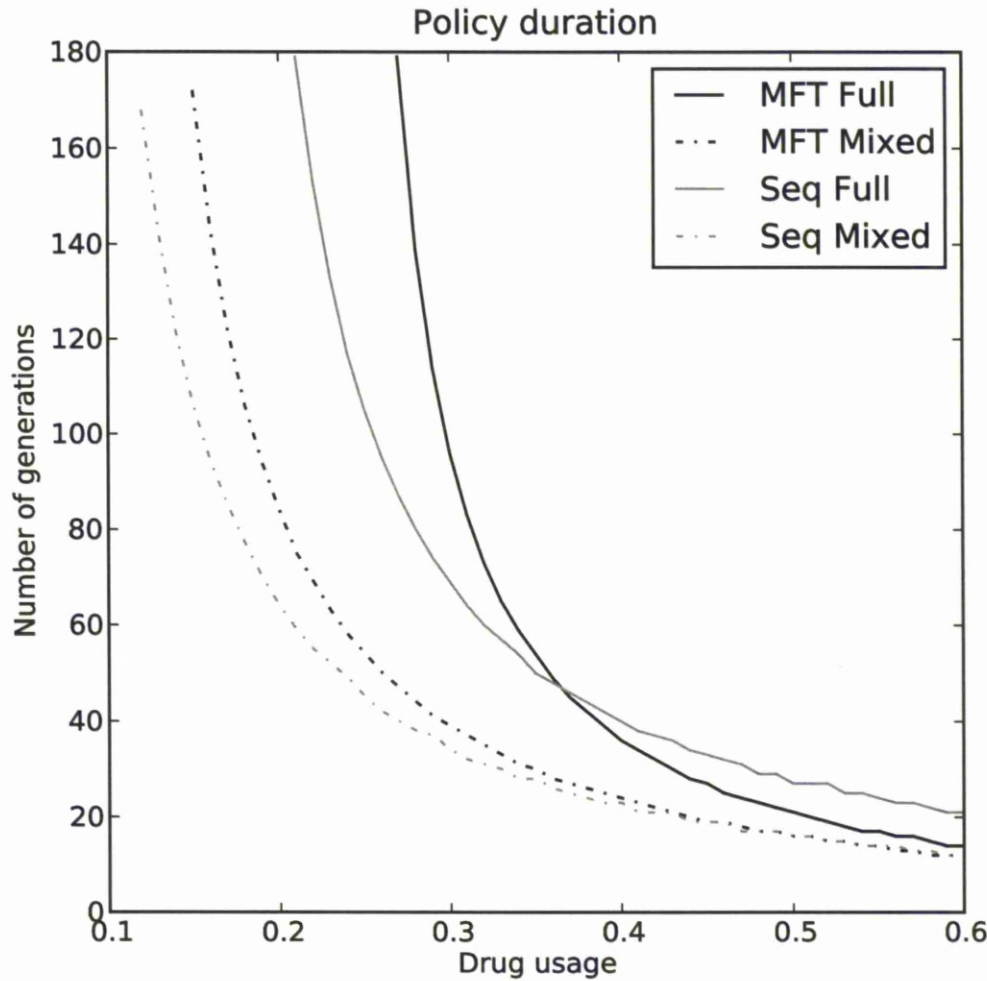


Figure 4.5: Impact of policy and epistasis on useful therapeutic life. We plot the policy duration on the Y-axis and the drug usage on the X-axis. The chart compares Full epistasis with the mixed model for MFT and sequential drug application assuming a fitness penalty of 10%. The useful therapeutic life always lasts longer in epidemiological settings requiring full epistasis rather than mixed model.

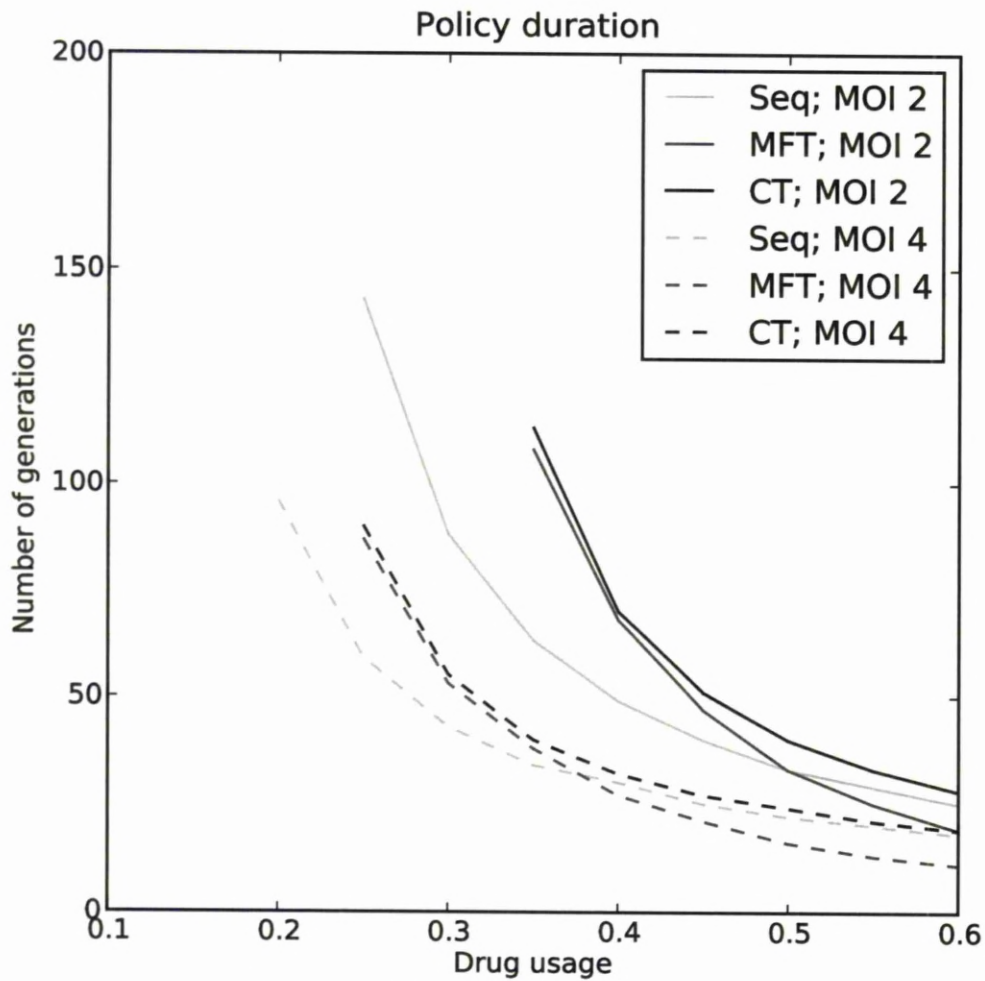


Figure 4.6: Impact of policy and MOI on useful therapeutic life. Three drugs are available (compared to two in Figure 1 of the main text) but all other parameters, abbreviations and simulation procedures are identical between the two figures. The lines start abruptly because for lower drug usage the policies last longer than the 200 generations simulated.

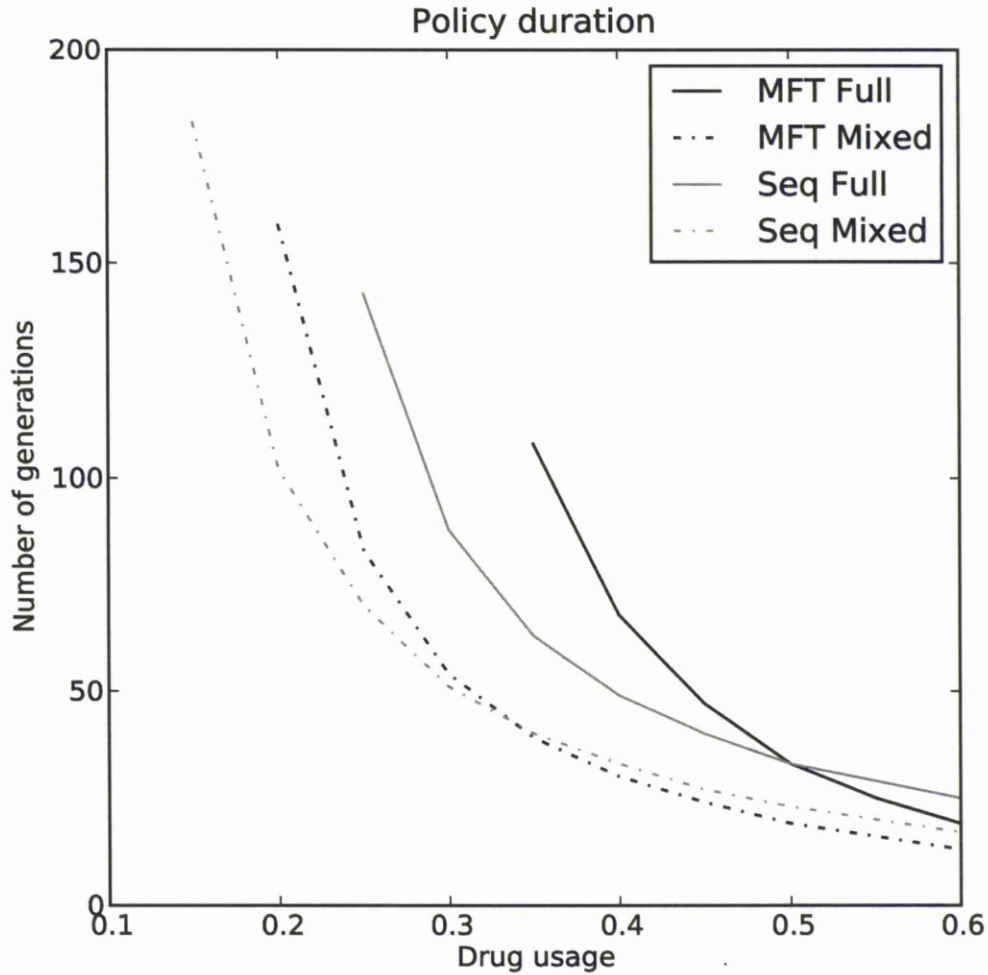


Figure 4.7: Impact of policy and epistasis on useful therapeutic life with 3 drugs. Three drugs are available (compared to two in Figure 2 of the main text) but all other parameters, abbreviations and simulation procedures are identical between the two figures.

Five

Evolutionary parasitology applied to control and elimination policies

Tiago Antao

Huijben et al. (2010) recently suggested that sub-curative antimalarial drug treatment of clinical cases might slow the spread of drug resistance through increased competition between drug resistant and sensitive infections. This was further discussed in letters to *Trends in Parasitology*. Goncalves and Paul (2011) raised important questions about the potential use of sub-curative treatment: What is the epidemiological and clinical impact of leaving a patient with circulating parasites? How is the correct drug sub-clearance level determined? Hastings (2011) added several ethical and operational arguments against such a proposal namely that patients treated with sub-curative drug levels may have repeated episodes of recurrent malaria and/or succumb to a secondary infection after being weakened by the primary malaria episode. However, the basic evolutionary argument remains unscathed, in untreated patients, competition between resistant and sensitive infections will benefit the latter as resistant mutations are expected to incur a fitness penalty. The question then arises: is there any way to exploit competition between infections with different resistance profiles infecting the same host?

Fortunately the parasite environment that Huijben et al. describes does already exist. Asymptomatic parasite carriers, mostly individuals that have acquired immunity due to repeated infection in high transmission areas, provide an environment where there is no drug pressure as long as they are not treated. In this environment sensitive parasites can out-compete resistant ones due to the absence of drug induced selection. Also the community-wide pressure against sensitive parasites is lower because less individuals need treatment so fewer individuals have low level of drugs, a decisive factor in the spread of tolerance and resistance (Hastings et al., 2002).

Elimination attempts, by definition, will have to target all infected humans (Targett and Greenwood, 2008), thus finding and treating asymptomatic carriers is a fundamental part of such efforts. This will jeopardise the reservoirs of sensitive parasites even if

the treatment used for elimination is different from the standard first-line therapy. In this best-case scenario, if the elimination attempt fails, the frequency of resistance will probably increase above the pre-intervention frequency as asymptomatic carriers, the safe-haven for sensitive parasites, are treated. The worst-case scenario, using the standard first-line therapy for elimination, will drive a dramatic increase in resistance should the elimination attempt not succeed. It is also likely that a failed elimination intervention will have worse consequences, for the spread of resistance, in high-transmission areas as a greater proportion of infections are asymptomatic due to semi-immunity conferred by repeated infection (Doolan et al., 2009).

On the other extreme of the policy spectrum (i.e. maximising the number of untreated individuals to increase the relative fitness of sensitive parasites), a more radical approach to sub-clearance treatment would be to provide only palliative care. This would presumably be applied only to uncomplicated malaria cases on a voluntary basis. Analogous situations arise in other infectious diseases like influenza (Fiore et al., 2011) where not all individuals are treated with anti-virals. Most unfortunately, malaria is not influenza, and the patient condition can deteriorate rapidly with complications like renal failure or even cerebral malaria which can manifest suddenly and have a high mortality rate (Mishra and Newton, 2009). Clearly, current adjunctive therapy cannot be seen as a palliative alternative to replace proper malaria treatment (John et al., 2010).

The underlying premise that drug resistant parasites are out-competed by sensitive parasites in non-treated environments is based in sound evolutionary genetic theory which is consistent, for instance, with the observation in Malawi (Laufer et al., 2006) where Chloroquine removal led to a rather rapid disappearance of CQ resistance presumably due to a fitness penalty of resistance mutations. Strategies to better exploit this effect are still not clear but rational public health policy should recognise that policies to eliminate and eradicate malaria might conflict with more modest strategies to pursue only control. Decision makers that embark on elimination policies should be reasonably sure that they will be able to meet the desired outcome as excessive optimism might lead to a serious control problem after the failure of well-meaning, but possibly disastrous elimination efforts. Therefore, proper risk-analysis encompassing many different factors ranging from donors' ability to maintain funding, transmission intensity, local geo-political factors of the intervention area and the ability to conduct proper surveillance and monitoring of an intervention among others is fundamental to assure that any elimination attempt will not develop into a difficult to control scenario. If the last man standing is indeed the most resistant (Maude et al., 2009), shouldn't at least be a Plan B in the case we indeed fail to cure him?

Part II

F_{ST} selection detection and discovering genes involved in drug resistance

Six

LOSITAN: A workbench to detect molecular adaptation based on a F_{ST} -outlier method

Tiago Antao, Ana Lopes, Ricardo J Lopes,
Albano Beja-Pereira and Gordon Luikart

Abstract

Background: Testing for selection is becoming one of the most important steps in the analysis of multilocus population genetics data sets. Existing applications are difficult to use, leaving many non-trivial, error-prone tasks to the user.

Results: Here we present LOSITAN, a selection detection workbench based on a well evaluated F_{ST} -outlier detection method. LOSITAN greatly facilitates correct approximation of model parameters (e.g., genome-wide average, neutral F_{ST}), provides data import and export functions, iterative contour smoothing and generation of graphics in a easy to use graphical user interface. LOSITAN is able to use modern multi-core processor architectures by locally parallelizing `fdist`, reducing computation time by half in current dual core machines and with almost linear performance gains in machines with more cores.

Conclusions: LOSITAN makes selection detection feasible to a much wider range of users, even for large population genomic datasets, by both providing an easy to use interface and essential functionality to complete the whole selection detection process.

6.1 Background

Understanding the contribution of selection and molecular adaptation in shaping genome wide variation is among the most exciting and widely researched problems with many applications ranging from human health to conservation of endangered species. Among the many selection detection strategies (Nielsen, 2005), F_{ST} outlier approaches are becoming

widely used (Beaumont, 2005; Vitalis and Couvet, 2001) because they are important not only for studying the genetic basis of adaptation but also for eliminating non-neutral outlier loci from data sets before computing most population genetic parameters (e.g., F_{ST} , N_m , N_e), that require neutral loci (Luikart et al., 2003). This is particularly important in a time where production of data sets with information from hundreds of loci is becoming fairly common.

One such F_{ST} method is described in Beaumont and Nichols (1996); Beaumont (2005) (but see also Cavalli-Sforza (1966) and Lewontin and Krakauer (1975)) and is implemented in the `fdist` program and can be used for any codominant genetic molecular markers including microsatellites, Single Nucleotide Polymorphisms (SNPs) and allozymes. This method evaluates the relationship between F_{ST} and H_e (expected heterozygosity) in an island model (Wright, 1931), describing the expected distribution of Wright's inbreeding coefficient F_{ST} vs. H_e under an island model of migration with neutral markers. This distribution is used to identify outlier loci that have excessively high or low F_{ST} compared to neutral expectations. Such outlier loci are candidates for being subject to selection.

Using `fdist` can be a challenging task for those not familiarized with command-line applications and requires a specific data format not used by other applications (Excoffier and Heckel, 2006). Furthermore, several independent runs are usually needed to tune parameters (e.g., determine the appropriate average F_{ST}) before a final execution is made in a process that is prone to human introduced mistakes. `Fdist`, not being one of the most computationally intensive programs available, can still take up to one hour for a single run (especially if smooth contours for confidence intervals are required), and, in most cases, multiple runs are needed for parameter tuning. Large population genomic datasets can take even longer. In this context, `fdist` requires experienced computer users, and its usage is error prone (e.g., by incorrectly converting data files or not approximating average F_{ST} appropriately).

6.2 Implementation

We designed LOSITAN (LOoking for Selection In a TANGled dataset), a selection detection workbench constructed around `fdist`. LOSITAN is a Java Web Start application coded mostly in Jython with a small part in Java, allowing direct execution from the web. LOSITAN provides the following features:

1. Easy to use interface (Figure 6.1), directly usable from the web.
2. Data import in Genepop (Raymond and Rousset, 1995) format.
3. Generation of graphics in several formats (PNG, SVG and PDF).

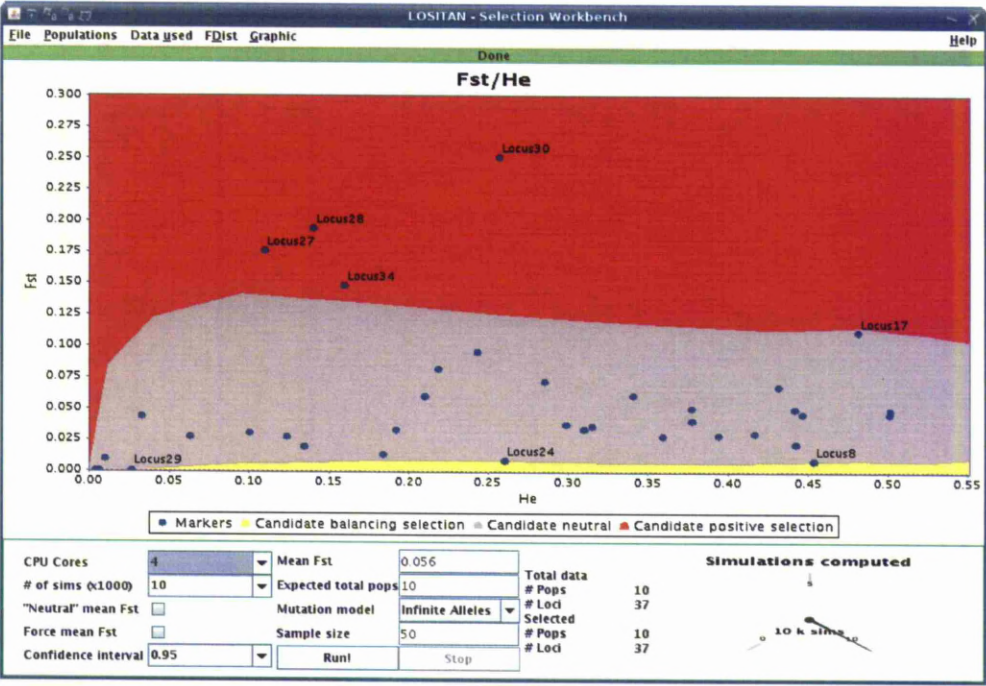


Figure 6.1: LOSITAN console: Screen shot showing run parameters (bottom panel) and a graphical output with the simulated confidence area for neutral loci (middle color band) with loci from the original empirical dataset represented as dots. Outliers are tagged with labels.

Graphics can be generated in several formats (covering both bitmap and vector format styles) and parametrized in many ways (from choosing colours to deciding which labels are printed, among others). A completely unedited example of a PNG output is presented on Figure 6.2.

4. Data export in a format suitable for import into statistical packages like R (R Development Core Team, 2007) or commonly used spreadsheet software.

In case the user desires to further analyze the data or have total flexibility in generating graphics, LOSITAN makes available both the confidence intervals computed and the F_{ST} s and heterozygosities for each locus.

A simple R script is supplied in order to facilitate loading the data into R. Loading in spreadsheet software is done simply by importing as a tab delimited file.

5. Choice of which populations and/or loci are studied.
6. Approximating mean neutral F_{ST} (in the real dataset) by removing potential selected loci.

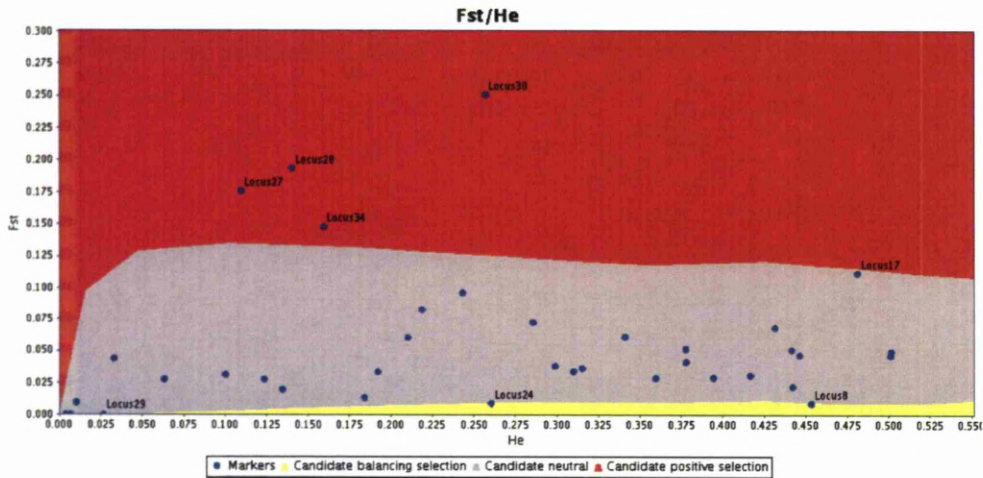


Figure 6.2: PNG output: Graphical output from LOSITAN in PNG format without any post-edition.

The initial mean dataset F_{ST} is often not neutral in the sense that (initially unknown) selected loci are often included in the computation. LOSITAN can optionally be run once to determine a first candidate subset of selected loci in order to remove them from the computation of the neutral F_{ST} . This value will be, in most cases, a better approximation of the neutral F_{ST} (Beaumont and Nichols, 1996). The procedure works as follows: LOSITAN is run a first time, using all loci to estimate the mean neutral F_{ST} . After the first run, all loci that are outside the desired confidence intervals (e.g. 99% CIs) are removed and the mean neutral F_{ST} is computed again using only putative neutral loci that were not removed. A second and final run of LOSITAN, using all loci, is then conducted using the last computed mean. This procedure lowers the bias on the estimation of the mean neutral F_{ST} by removing the most extreme loci from the estimation. Naturally all loci will be present in the last run and will have their estimated selection status reported.

7. Approximating average simulated F_{ST} to the average value found in the real dataset even when the experimental conditions are far from the ones where the theoretical formula $F_{ST} = \frac{1}{4Nm+1}$ holds (e.g. low number of demes or the usage of the stepwise mutation model, common in microsatellite markers).

To be able to (optionally) approximate the average F_{ST} in conditions far from the theoretical optimum, LOSITAN starts by running `fdist` for 10,000 realizations using the theoretical value, calculating the average simulated F_{ST} , if the value is too far from the real average F_{ST} , LOSITAN uses a bisection approximation algorithm running 10,000 realizations for every tentative bisection point. The algorithm

works by iteratively slicing the interval of possible F_{ST} values (i.e., between 0 and 1) in half at each iteration and choosing the mean of the bounds on each iteration (with the exception of the first iteration where one of the extremes is chosen). An example is provided to make the approach clearer:

In a certain demographic scenario we want to simulate a neutral F_{ST} of 0.08. The algorithm starts by trying 0.08. If the result is higher than desired then 0.0 will be tried (creating an absolute lower bound limit), after that 0.04 $(0.0 + 0.08) / 2$ will be tried, if the result is too low, 0.06 will be used next (i.e. $(0.04 + 0.08) / 2$), the process repeats until the error margin is acceptable.

In practical terms the method was able to converge to the desired value in all cases tested (a completely trivial bisection approach is not possible as the method for computing F_{ST} is stochastic and results might vary for the same input conditions).

8. Iterative smoothing of confidence interval contours.

Contour smoothing is achieved by running `fdist` an extra 5,000 realizations. The user can request smoothing an unlimited number of times until the result is deemed satisfactory.

9. Ability to use multiple CPU cores and processors when running `fdist`.

To be able to use multiple cores, `LOSITAN` divides the number of desired simulation repeats among all available cores (although the application detects the number of existing cores, the user is able to change the number of simultaneous concurrent processes), this is possible because `fdist` simulation runs are independent, thus making parallelization a simple task. Tests show a near linear relationship between the number of cores used and performance gains, an existing 5-10% penalty is due mainly to joining the partial results together. `LOSITAN`, although being directly executable from the web is a client-side application and all computational intensive operations occur on the user computer and not on the server.

10. Automatic and transparent download of the latest version of `fdist`.

We maintain the latest version of the `fdist` application on the server, which is downloaded transparently by the client application whenever there is a new version. At the time of this writing the supported version is `fdist2`.

The interface includes tips for all the less obvious parameters and enforces constraints for all the user inputs which the system can infer are not correct.

6.3 Results and Discussion

In a beta test release to users the feedback was generally very positive stressing essentially that the application is easy to use, allows to easily input and output data and

deal with non-trivial parameter determination like calculating neutral F_{ST} . Most importantly it made users aware of issues in data analysis that they were not aware of. For example, users were not aware of how to estimate the genome wide average neutral F_{ST} from their empirical data set by removing one or a few strong outlier loci, and the recomputing the average F_{ST} . Although LOSITAN helps avoid many pitfalls involved with using F_{ST} -outlier approaches in general, it is not able to solve fundamental issues regarding these approaches, for instance the non-linear behavior of $F_{ST} = \frac{1}{4Nm+1}$ when F_{ST} approaches zero can make it difficult to detect low F_{ST} -outliers especially when selection is not strong. As such an easy to use application should not be seen by users as a excuse to avoid critical reasoning around the the whole selection detection process. Feedback from users also allows to chart possible future work, like supporting dominant markers or supporting other selection detection approaches like Vitalis and Couvet (2001).

The most important and recurrent request from users pertains the support of large amounts of markers (mostly SNPs) that are being generated by next generation sequencing technologies. When version 1 of LOSITAN was released, typical datasets included less than 100 markers whereas now tens of thousands of markers are possible and even hundreds of thousands of SNPs are becoming common. LOSITAN version 1 would load all the genetic information into memory, thus any dataset bigger than available memory would not be readable by the application. LOSITAN version 2, uses an iterative parser which reads one individual at time and updates summary information making the application usable with virtually unlimited markers. LOSITAN version 2 was tested with real datasets of up to 25,000 markers. While LOSITAN has currently no limit, `fdist2` is limited to around 50,000 markers and, therefore, there is still a practical limit. If future users require more than 50,000 markers, changing the hard-coded limit of `fdist2` should be feasible.

Our solution to use all the available computing power on new multi-core hardware is an example of an “embarrassingly simple parallel” computation approach. We contend that having a simple approach is a good principle: The point in this application is to make all computational power available to the users and not to develop new concurrent algorithms. A simple, highly efficient, elegant and less bug-prone approach is what responds to the users needs, as the objective of this work is not to develop new algorithms, but to use them.

6.4 Conclusions

LOSITAN is built along the principles exposed in Kumar and Dudley (2007), namely that intuitiveness and user empowerment should be fundamental guidelines for software construction targeting biologists. This is done, not only by supplying an easy to use web interface for an, otherwise, hard to use application, but also allowing the use of widely

utilized population genetic data formats, automating the tuning of nuisance parameters and lowering the computational costs on modern hardware. In addition, strong emphasis is put on trying to avoid errors on the usage of the software either by both enforcing constraints and giving suggestions on less obvious features. This will lower the barriers to usage of the underlying application, allowing for a wider user base which will be able to concentrate more on the biological problems and less on unnecessary application complexity.

We are in the dawn of the era of multi-core computing. The vast majority of existing software cannot make use of the extra computational power made available on new machines. Our approach, based on partitioning a computational intensive task into smaller ones, can be used to leverage the extra computational power even without changing existing code on applications which can be broken into smaller independent running units. This partitioning approach can be performed in some cases by users on existing software or by programmers in new applications that take advantage of multiple cores. With the current trend of supplying many more cores with new computers, strategies like the one presented here will be mandatory in order to take full advantage of all the existing processing power. LOSITAN is one of the first of many applications to explore the multi-core programming paradigm.

Future planned developments will include addition of other F-outlier methods and simulation facilities for explore the effects of different demographic scenarios on F_{ST} variance and the detection of outliers.

All the code to handle GenePop and fdist file formats and applications was also donated to the Biopython project and is publicly available starting from version 1.44.

6.5 Availability and requirements

Project name LOSITAN

Project home page <http://popgen.eu/soft/lositan>. Development site: <http://code.google.com/p/lositan/>

Operating systems Platform independent

Programming language Java and Jython

Other requirements Browser with JavaWebStart to run over the internet (software can be run locally).

Windows: At least Windows 2000 and Java 1.6.

Mac OS X: 10.4 (Tiger) and Java 1.5 (Most current 10.4 installations will require a freely available Java update).

Linux: Java 1.6 and the free GNU C compiler.

License GNU GPL

Any restrictions to use by non-academics None

Seven

Mcheza: A workbench to detect selection using dominant markers

Tiago Antao and Mark A. Beaumont
joint-first authorship

Abstract

Motivation: Dominant markers (DArTs and AFLPs) are commonly used for genetic analysis in the fields of evolutionary genetics, ecology and conservation of genetic resources. The recent prominence of these markers has coincided with renewed interest in detecting the effects of local selection and adaptation at the level of the genome.

Results: We present Mcheza, an application for detecting loci under selection based on a well evaluated F_{ST} -outlier method. The application allows robust estimates to be made of model parameters (e.g., genome-wide average, neutral F_{ST}), provides data import and export functions, iterative contour smoothing and generation of graphics in an easy to use graphical user interface with a computation engine that supports multi-core processors for enhanced performance. Mcheza also provides functionality to mitigate common analytical errors when scanning for loci under selection.

Availability: Mcheza is freely available under GPL version 3 from <http://popgen.eu/soft/mcheza>.

7.1 Introduction

Non-specific amplification methods such as Diversity Arrays Technology (DArT) markers and amplified fragment length polymorphism (AFLP), are commonly used for analysis of within-species variation because they allow the rapid acquisition of substantial genetic information, at relatively low cost. Although other alternative sequencing techniques have since been developed, DArTs and AFLPs are still widely used in the fields of evolutionary genetics, ecology and conservation. One of the most important applications

of these dominant markers is in detecting the effects of selection and local adaptation at the level of the genome, in areas ranging from parasitology to conservation genetics.

There are two current approaches to detect selection: “classical” F_{ST} -outlier approaches (reviewed in Storz, 2005), based on the distribution of summary statistics; and those based on likelihood (such as Beaumont and Balding, 2004; Foll and Gaggiotti, 2008). The original F_{ST} -outlier methods do not account for dominant markers such as DArTs or AFLPs. These markers have two phenotypes detectable at each locus: one allele (the plus-allele) is amplifiable, whereas the other (the null-allele) is not. Heterozygous genotypes cannot be directly distinguished from homozygotes making the estimation of allele frequencies non-trivial.

A widely used F_{ST} -outlier method (Pérez-Figueroa et al., 2010) for detecting selection with dominant markers is implemented in the package DFDIST. DFDIST is a modification of the FDIST program (Beaumont and Nichols, 1996) to allow for dominant markers, and implements the method of Zhivotovsky (1999) to estimate allele frequencies. Briefly, coalescent simulations are used to generate a null sampling distribution of estimates of F_{ST} based upon neutral expectations. The performance of the method has been examined, using data simulated with known levels of selection, in Caballero et al. (2008) and Pérez-Figueroa et al. (2010). DFDIST has a complicated text-based interface that makes it difficult to use correctly. For example the tuning of parameters for the coalescent simulations is non-trivial, and the input dataset has to be formatted in a non-standard way. We describe Mcheza, a new application based on DFDIST, with a graphical user interface allowing easier configuration of some non-trivial parameters.

7.2 Software implementation

The Mcheza architecture is composed of two parts: the front-end implemented in Jython and the DFDIST back-end implemented in C. The front-end provides an interface similar to LOSITAN (Antao et al., 2008) (A selection workbench based on the analogous method for co-dominant markers). The interface provides the following functionality on top of DFDIST:

1. Estimation of the mean neutral F_{ST} , while taking into account loci that might be under selection. While DFDIST requires an estimate of the neutral F_{ST} , an empirical dataset will probably include loci under selection. Mcheza provides a mechanism similar to that in LOSITAN for estimating the neutral F_{ST} based on removal of loci that are potentially under selection.
2. An improved method for ensuring that the simulated distribution of F_{ST} has a mean that is close to the required value. DFDIST is only capable of providing a reliable approximation when close to theoretical conditions (i.e. when simulating

a large number of populations). The Mcheza interface provides a correction that accurately approximates F_{ST} even when the number of demes is very low.

3. Mcheza provides additional features in comparison with LOSITAN by supporting very large datasets: while LOSITAN is only able to support hundreds of loci and hundreds of individuals, Mcheza has been tested using real datasets with 25,000 loci. Support for very large datasets is, as expected, computationally more intensive.
4. Mcheza also introduces support for multi-test correction based on false discovery rates (FDR) (Benjamini and Hochberg, 1995), as implemented in Chiurugwi et al. (2010). Without such a correction there is a danger in over-estimating the proportion of loci that are under selection (Beaumont, 2008; Pérez-Figueroa et al., 2010).
5. A multi-core aware version of DFDIST with computational performance gains that are near linear with the number of cores.
6. An easy-to-use interface including the ability to import data in the standard Genepop format (Rousset, 2008), generation of charts, export in standard formats including R (R Development Core Team, 2007) and spreadsheets, iterative smoothing of confidence contours, choice of population and loci among other features.

The application, based on the Java Web Start technology, requires only a browser with a modern version of Java installed (on Linux the GNU C compiler is also required). The Java code will detect the operating system and choose the correct DFDIST implementation.

The use of the Jython programming language allows the use of Biopython (Cock et al., 2009) which provides a parser for Genepop files. Our Python code to interact with DFDIST is incorporated in the Biopython population genetics module allowing for bioinformatics programmers to directly interface with the DFDIST core using the Python programming language.

7.3 Discussion

A fundamental consideration in the design of Mcheza is supporting the user by correctly computing important non-trivial parameters that are needed to properly calculate candidate loci for selection. Erroneous usage of population genetics applications can easily produce results that seem correct but are, in effect wrong.

While Mcheza tries to minimise usage errors, the user should be aware of potential limits of the underlying method. Potential users are advised to read Caballero et al.

(2008), Pérez-Figueroa et al. (2010) and Excoffier et al. (2009) where several scenarios where DFDIST is less applicable are clearly explained. In particular, the Bayesian method of Foll and Gaggiotti (2008), implemented in the program *BayeScan* is an important alternative, with which results should be compared. By improving the estimation of the neutral mean F_{ST} when the number of demes is low, Mcheza addresses situations where DFDIST is known to perform less well (Pérez-Figueroa et al., 2010).

In summary Mcheza tries to provide an intuitive interface, which includes intelligent suggestions to the user with regards to correct usage of software, while enforcing model constraints and providing necessary corrections (e.g. FDR support). It is hoped that this approach will lower barriers to its use, allowing researchers to concentrate more on the biological problems (including the theoretical assumptions and limitations of underlying models) and less on unnecessary software complexity.

Eight

interPopula: a Python API to access the HapMap Project dataset

Tiago Antao

Abstract

Background: The HapMap project is a publicly available catalogue of common genetic variants that occur in humans, currently including several million SNPs across 1115 individuals spanning 11 different populations. This important database does not provide any programmatic access to the dataset, furthermore no standard relational database interface is provided.

Results: interPopula is a Python API to access the HapMap dataset. interPopula provides integration facilities with both the Python ecology of software (e.g. Biopython and matplotlib) and other relevant human population datasets (e.g. Ensembl gene annotation and UCSC Known Genes). A set of guidelines and code examples to address possible inconsistencies across heterogeneous data sources is also provided.

Conclusions: interPopula straightforward and flexible Python API facilitates the construction of scripts and applications that require access to the HapMap dataset.

8.1 Background

The HapMap project (International HapMap Consortium, 2007) (<http://hapmap.ncbi.nlm.nih.gov/>) is an effort to identify and catalogue genetic similarities and differences in humans. The project makes information available on single nucleotide polymorphisms (SNPs), and it more recently added information on copy number variation (CNV). HapMap phase 3 includes data on 1115 individuals (around 1.5 million SNPs per individual) spanning 11 populations while phase 2 included only 4 populations (270 individuals) but more than 3.5 million SNPs per individual. This dataset can be useful in a multitude of situations from finding genes that affect human health to evolutionary

research about the human species or for genome-wide association studies. All of the information generated is released into the public domain and can be downloaded with minimal constraints.

The HapMap project provides access to the data in bulk form (via FTP download), a web interface (Thorisson et al., 2005) which includes a genome browser (Stein et al., 2002) and the data mining application HapMart based on Biomart (Smedley et al., 2009). Programmatic and relational database interfaces are not offered though some API support is implemented by external parties such as a generic Perl API for variation datasets in Ensembl (Rios et al., 2010), BioPerl's Bio::PopGen module (Stajich et al., 2002) or the GTools package (Carey et al., 2007) for R/Bioconductor. Most existing libraries support only a subset of features (e.g. parsing of HapMap file formats or creating a local database) making the construction of scripts and applications more complex as basic data manipulation functionality must be built as least partially. Furthermore, there is no known Python library supporting HapMap data.

8.2 Implementation

interPopula provides a Python API to access the HapMap dataset. Interfaces to all HapMap phases are supported including phase 2 data with fewer populations but more SNPs genotyped per individual and phase 3 covering more populations. interPopula provides access to frequency, genotype, linkage disequilibrium and phasing datasets. The recent CNV dataset is also supported along with family relationships for the 5 populations where sampling was performed for family trios (mother, father and one offspring).

Support for annotation information that is commonly needed to process HapMap data is also provided through an API to both the UCSC Known Genes dataset (Hsu et al., 2006) from the UCSC genome browser database (Rhead et al., 2010) and the Ensembl gene annotation database (Curwen et al., 2004).

The API was constructed according to the following design guidelines:

1. The API is straightforward and self-contained. The core API requires only a Python interpreter, has no extra dependencies and minimal administrative overhead.
2. Downloaded data is stored on an SQL database for faster access. All data is stored using SQLite (SQLite Development team, 2010) which is natively supported in Python thus lowering the maintenance costs of the system. interPopula can also be connected to enterprise-grade databases which support multiple users, concurrent usage and large datasets for which the standard sqlite backend might not be enough (a PostgreSQL example is provided).

3. Data management (i.e. downloading from the HapMap site and local database construction) is fully automated: the required data subset is downloaded on demand only once and stored locally, reducing the load on both the client and server.
4. While SQL interfaces are made available from both the UCSC and Ensembl projects for their annotation databases, interPopula uses the same implementation strategy for the HapMap dataset: files are intelligently downloaded and locally stored. This provides a consistent interface to these two datasets which provide important annotation information frequently used to process HapMap data.
5. The framework is extensible and designed to be easily integrated with other Python tools and external databases. The web site provides several examples of integration with standard tools used in Python for bioinformatics such as Biopython (Cock et al., 2009), NumPy (Oliphant, 2006) and matplotlib (Hunter, 2007).
6. Integration with Biopython allows for access to the Entrez SNP database and the population genetics tools supported by Biopython such as Genepop (Rousset, 2008) allowing automated analysis of datasets.
7. Facilities to export HapMap data to Genepop format are provided enabling (non-automated) analysis of the HapMap dataset with the plethora of population genetics software which support this format. Data export can also be used to initialize population genetics simulators like the Python-based simuPOP (Peng and Kimmel, 2005) allowing computational simulations to be initialised with real datasets.
8. A large set of scripts is included, serving both as utilities to analyse the data, as well as examples of database and external tool integration. Currently we provide examples of integration with Entrez databases (nucleotide and SNP), the Genepop population genetics suite and charting libraries.
9. A set of guidelines and scripts was developed in order to facilitate a consistent view across heterogeneous databases. HapMap, Ensembl, UCSC Known Gene and the Entrez databases might not be fully consistent among themselves and, if care is not taken, database integration efforts might lead to erroneous results. The main pitfall is the usage of different NCBI reference builds across different databases, most notably HapMap is still based on build 36 whereas other databases either support multiple builds or only the most recent build 37.
10. A robust open-source software development process is put in place: a full public web based platform (hosted on Launchpad) is used to maintain the code infrastructure and unit tests approach 100% coverage.

8.3 Results

interPopula can be used to create a wide range of applications and scripts based on the HapMap dataset. The most commonly expected usage pattern will be for genome wide association studies, though the example presented here will be of a different nature.

As an example of usage, we present a population comparison of all the genotyped SNPs for a gene. We will plot the F_{ST} statistic for all Lactase SNPs between two HapMap populations: Utah residents with Northern and Western European ancestry (CEU) and Yoruban in Ibadan, Nigeria (YRI). These populations are known to differ in their tolerance to lactose (Tishkoff et al., 2007). This example uses genotype information from HapMap and also demonstrates the integration facilities with UCSC Known Genes (to retrieve gene position and exon data), matplotlib (used for plotting), Biopython and Genepop (used to calculate F_{ST}).

This example, which is quite complex in terms of integration between several databases and tools can be broken down into the following steps:

1. Load the Known Genes database. The version pertaining build 36 should be loaded to assure consistency with HapMap.
2. Determine relevant information about Lactase from Known Genes. The following information is needed: The chromosome on which it is located, the start and end positions in the chromosome and all exon positions.
3. Load HapMap genotype information for the CEU and YRI populations for the relevant chromosome.
4. Retrieve all the HapMap SNP ids between the start and end positions in the chromosome.
5. Export a Genepop formatted file with two populations including all HapMap SNPs for Lactase.
6. Call the Genepop application via Biopython to calculate the F_{ST} for all markers.
7. Plot the calculated F_{ST} s along with the exon positions.

The result of this example is shown in Figure 8.1. The X-axis reports the position along chromosome 2, F_{ST} in on the Y-axis, the dots represent the F_{ST} values for existing SNPs on the HapMap database and the red boxes are the exon positions (17 in the case of Lactase). Interpreting the results of this specific application of interPopula is beyond the scope of this manuscript but at least two different interpretations are possible: (i) SNPs where F_{ST} is above approximately 0.45 are candidates for positive selection (as around 95% of F_{ST} values for humans are below 0.45 (Akey et al., 2002)) or (ii) the F_{ST} statistic is noisy when applied to a single marker (Kelley et al., 2006) .

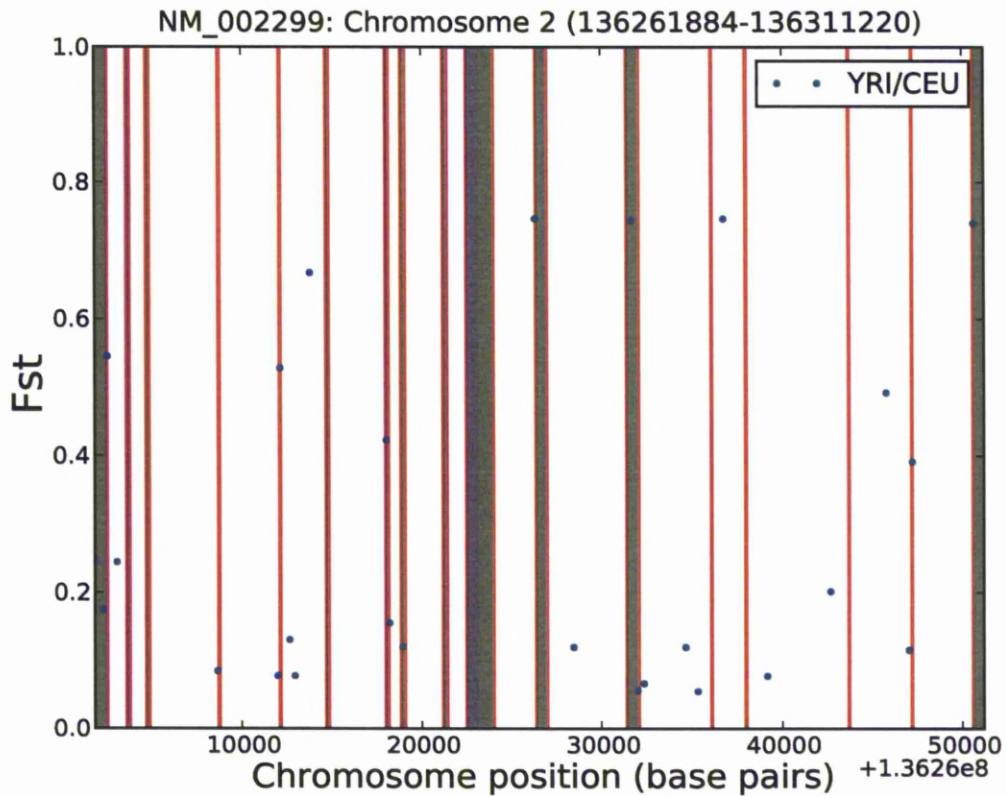


Figure 8.1: F_{ST} for Lactase between 2 HapMap populations: F_{ST} between CEU and YRI populations for all Lactase SNPs on the HapMap database. The X-axis reports the position on chromosome 2 (the value on the lower right is the absolute offset from the beginning of the chromosome), the Y-axis the F_{ST} value. The dots represent the F_{ST} values for existing SNPs on the HapMap database. The red boxes represent exon positions. To construct this chart HapMap frequency data and the UCSC Known Genes database were consulted. Biopython and Genepop were used to compute the F_{ST} statistic.

The above example was constructed using the UCSC Known Genes database but the programmer can alternatively use the Ensembl gene annotation database instead.

This example (script `IGFstGene.py` in the distribution), along with more than 20 others including data export, connection to enterprise-grade databases, analysis of the distribution of the number of exons per gene, the distribution of genes per chromosome are made available with `interPopula`.

In order to illustrate `interPopula`'s basic API, Figure 8.2 shows a commented script which provides useful functionality. In this example the HapMap frequency database is consulted to report the frequency of both alleles for each SNP within a certain chromosome interval. The code example is less than one page in length and there are only 4 API calls to achieve the complete functionality. This is one case illustrating the ease

of use of the API. All scripts provided with interPopula are documented to the level of the example presented and automated documentation covering the full API is extracted from the source using epydoc (<http://epydoc.sourceforge.net/>).

The part of the API devoted to both UCSC Known Genes and the Ensembl gene annotation database can be used stand-alone to access both databases, i.e., it can be used for application and scripts that have no relationship with the HapMap data. interPopula's UCSC and Ensembl APIs can be used to access also non-human data as genome annotations are available for other species. This is especially useful with the Ensembl dataset as it makes available gene annotation information for many other species. Users should note that the quality of the datasets for other species varies as more effort is put in the curation of human data (e.g. while for humans the chromosome information is normally the chromosome number, for cats – *Felis catus* – it is mostly scaffold data). Stand-alone example script examples are provided for both datasets.

Future development efforts for interPopula will focus on supporting large datasets. As sequencing costs continue to decrease and the sequencing of complete genomes becomes commonplace it is clear that the backend infrastructure will have to be redesigned to support the large amounts of data generated by such efforts. In this context, supporting the 1000 genomes project (The 1000 genomes project team, 2010) is a natural extension for interPopula as many of the samples used in this project come from the HapMap dataset. While the API for UCSC and Ensembl extensions provides access to other species data, the main focus of interPopula will remain providing robust and well-maintained APIs for publicly available human genomic datasets which lack a standardized Python API or relational interface.

8.4 Conclusions

interPopula is a flexible, straightforward Python API to the HapMap project. It strives to integrate with both common Python bioinformatics and scientific libraries and other genomic databases that are commonly used in conjunction with the HapMap dataset. interPopula makes HapMap data processing possible inside Python, thus opening the possibility for the development of a plethora of interesting applications and scripts that make use of this important resource for human population genomics studies.

8.5 Availability and requirements

Project name interPopula

Project home page <http://popgen.eu/soft/interPop/>. Development site: <https://launchpad.net/interpopula>

Operating systems Platform independent

Programming language Python

Other requirements Optionally NumPy, Biopython, Genepop and matplotlib

License GNU GPL version 3

Any restrictions to use by non-academics None

```

from interPopula import Config
from interPopula.HapMap.Frequency import Frequency

#configuration directory
Config.dataDir = "."

#Han Chinese Lactase information
pop = "CHB"
chr = 2
startChr = 136261855
endChr = 136311220

#Lets get the Frequency information
freqDB = Frequency("2010-05_phaseIII")

# We require a chromosome and population
freqDB.requireChrPop(chr, pop)

# We need to get the RSIDs for the interval
RSs = freqDB.getRSsForInterval(chr, startChr, endChr)

print "rsid allele1 freqAllele1 allele2 freqAllele2"
for rs in RSs:
    #We get frequency information
    freqSNP = freqDB.getPopSNPs(pop, rs)
    #a1 retrieves allele 1 (A,C,T,G), a2 does the same for 2
    a1 = freqSNP[5]
    a2 = freqSNP[6]
    #frequency of a1 homozygotes
    a1a1 = freqSNP[7]
    #frequency of a2 homozygotes
    a2a2 = freqSNP[8]
    #frequency of heterozygotes
    a1a2 = freqSNP[9]
    #gets the frequency of allele 1
    fa1 = (2.0*a1a1+a1a2)/(2*a1a1 + 2*a2a2 + 2*a1a2)
    print rs, a1, fa1, a2, 1 - fa1

#example output
#rsid allele1 freqAllele1 allele2 freqAllele2)
#730005 C 0.727941176471 T 0.272058823529
#872151 C 0.658088235294 T 0.341911764706
#1042712 C 0.36496350365 G 0.63503649635
#[...]

```

Figure 8.2: Example code to print the frequency of HapMap SNPs: This example describes how to consult the HapMap frequency database to retrieve the allele frequencies for a set of SNPs in section of a chromosome.

Nine

Detecting F_{st} -outliers and selection requires genotyping multiple SNPs per gene: lessons from empirical data

Tiago Antao and Gordon Luikart

Abstract

Single nucleotide polymorphisms (SNPs) are increasingly used to identify genes under selection using F_{ST} -outlier tests. However, researchers often genotype only a few SNPs per gene. We quantify the sensitivity of using only a few SNPs in a gene to identify high F_{ST} -outlier genes using large empirical data sets from humans. We find that genotyping less than four random SNPs generally gives low power (<80%) to detect genes that are known to be under directional selection, because of high variation in F_{ST} among SNPs. These results held even for genes with the largest proportion of high- F_{ST} SNPs. For example, alleles in the lactose tolerance gene have been under strong selection in humans in Northern Europe. However, only 15 of 61 SNPs across the gene show excessively high F_{ST} (>0.45). Inferring haplotypes using 4–8 random SNPs seldom increased the power to detect high F_{ST} -outliers. We recommend genotyping >3 SNPs per gene to have a reasonable probability of identifying high F_{ST} -outlier genes. Genotyping >3 SNPs per gene will not substantially increase the false positive rate. Importantly, our results suggest that common SNP genotyping strategies (e.g. SNP chips, RADs, exon capture) often have too few SNPs per gene region to reliably detect selection.

Keywords: Statistical power, F_{ST} , HapMap, selection, molecular adaptation, single nucleotide polymorphisms, candidate genes.

9.1 Introduction

“The problem is that our knowledge about false negatives is even more rudimentary than about false positives. For panmictic populations, the power of

many tests to detect selection is known to be rather low. For a structured population, this information is basically missing."

Hermisson (2009)

Many genome scans and candidate gene studies that test for natural selection use only one or a few SNPs per locus (along with tests for F_{ST} -outliers) because it is costly to discover and genotype many SNPs per locus (Narum and Hess, 2011; Bourret et al., 2011). Unfortunately, high variation in F_{ST} and weak linkage disequilibrium among SNPs in a gene could lead to low power to detect F_{ST} -outliers, even following a strong selective sweep driving a certain haplotype or SNP(s) to high frequency in a local population.

Computational simulations are the most common approach to research the performance of methods to detect selection (see e.g. Beaumont and Balding, 2004; Vitalis et al., 2003). Use of computational simulations, instead of empirical data, is helpful because for most genomes and populations little is known about which genes are under selection, where complete understanding of selection exists for simulated populations. Unfortunately computational simulations have limitations such as the use of simplifying assumptions about demography, genomic structure, recombination rates, and patterns and strength of selection.

For some species (e.g. humans or pathogens such as *Plasmodium falciparum*) there have been extensive genome and population studies and several genes are known to be under selection. In some cases the mode of action of the gene and the mutations that affect protein function are also known. If the function, relevant mutations and/or mode of action of a gene are known then it becomes possible to evaluate a selection detection method against empirical data: If we know that a certain locus is under directional selection, what is the F_{ST} for individual SNPs and haplotypes? Does F_{ST} indicate directional selection? How many SNPs are needed to detect a signal of directional selection?

If large genome-wide datasets are available with reliable annotation, we can also search for all genes that have the largest proportion of SNPs with high F_{ST} and compute the probability of detecting such a gene when only using a small number of SNPs. We can then answer the question: How many SNPs must we genotype in order to have high power (above 80%) to detect a high F_{ST} locus?

Many species have much less information available than humans, at several levels: (i) There is no assembled genome reference sequence; (ii) gene function is mostly unknown or (iii) most studies lack the resources to genotype thousands of SNPs in thousands of genes. Here we use the information from the human HapMap project (International HapMap Consortium, 2007) which includes millions of genotyped SNPs and also available information about several human genes known to be under directional selection (e.g. involved in skin pigmentation, lactase persistence or protection against malaria).

We then “sample” from the HapMap dataset using only a few SNPs per gene (from one to eight, typical of studies in many other species) and ignoring any information about gene function and location. We then compare our findings by searching “blindly” with the known information about existing genes. Can we find genes that are known to be under selection using less information (few SNPs and no information about gene function)? This question is important for non-model species, and can also be useful with model taxa, as studies genotyping small numbers of markers are still common.

The main question that we address in this paper can be broadly expressed as: “If there is a gene under selection, (or with many high F_{ST} SNPs) how many SNPs must we genotype in order to detect it?” We are thus concerned with lowering the number of “false negatives”, a rarely addressed question with F_{ST} -outlier approaches (Hermisson, 2009). Here we evaluate and propose genotyping strategies to help researchers detect directional selection and avoid “false negatives”. We also briefly address the impact of our proposals on the false positive rate (i.e. erroneously concluding a low- F_{ST} locus (e.g. neutral locus) has exceptionally high- F_{ST}). This study has applications to current research in model and non-model species, namely with:

1. Limited genome scans where approximately 100–1000 SNPs are sampled to detect selection in genes (Narum and Hess, 2011; Bourret et al., 2011).
2. Candidate gene studies where tens-to-hundreds of genes are genotyped (using SNPs discovered or genotyped with techniques such as exon capture and next generation sequencing (Cosart et al., In press)).

9.2 Methods

We used HapMap Project Phase II+III data files (release 28) (International HapMap Consortium, 2007; Altshuler et al., 2010) for SNP analysis. Two populations were used: Utah residents with ancestry from northern and western Europe (CEU) and Yoruban in Ibadan, Nigeria (YRI). We removed offspring from the CEU and YRI samples and analysed only the data from 60 unrelated parents for each population. We used haplotype phase estimation performed using parent-offspring trios and imputation (Howie et al., 2009) by the HapMap consortium. In total, the numbers of polymorphic SNPs analysed for each population were 279,448 and 155,772 for chromosomes 2 and 12 respectively plus all SNPs for studied genes from other chromosomes (Table 9.1). All HapMap data used is based on NCBI genome build #36, thus all physical positions reported will be related to that build. To obtain human gene information (genomic location, number, position and size of exons) we used the UCSC Known Genes database (Hsu et al., 2006) (if different gene annotations were available for the same gene, we use the one which the largest gene size). F_{ST} was calculated according to Weir and Cockerham (1984) as implemented in Genepop (Rousset, 2008). F_{ST} was calculated for individual SNPs

and for haplotypes across each gene (blocks). Automated downloading and processing of HapMap data was done using *interPopula* (Antao, 2010). For automated bulk calculation of F_{ST} we used *Biopython* (Cock et al., 2009) to control *Genepop* and parse its output.

We studied the F_{ST} for individual SNPs and haplotypes for 5 genes that are known to have been under directional selection (see Tables 9.1 and 9.2). For chromosome 2, we also identified the 5 genes (with >20 SNPs genotyped) with the highest proportion of F_{ST} SNPs above 0.45 (hereby termed high- F_{ST} SNPs) and again studied individual SNP and haplotype F_{ST} . We computed for all genes the mean, median and maximum SNP F_{ST} . We also computed the probability of finding a SNP with F_{ST} above 0.45 when genotyping sets of 2, 4 or 8 SNPs randomly sampled (without replacement within a sampled set).

We used the haplotype information made available by the HapMap project and computed haplotypes with 2, 4 and 8 SNPs. SNPs were randomly chosen not using any existing information about tag SNPs (as tag information is normally not available for non-model species). We then also computed F_{ST} using haplotype reconstructions for each set of 2, 4 and 8 randomly sampled sets of SNPs.

To quantify the false positive rate, we computed the probability of detecting at least one high- F_{ST} SNP using 1, 2, 4 and 8 SNPs in the following cases: (i) random sample of SNPs across a whole chromosome (using chromosomes 2 and 12), (ii) random samples of SNPs inside a 150 kb interval (from chromosome 2 or 12); and (iii) random samples of SNPs inside genes retrieved from the UCSC Known Gene database for chromosomes 2 and 12.

9.3 Results

Among the 5 human candidate adaptive genes under directional selection that we studied, all have at least one SNP with F_{ST} above 0.45 (Figure 9.1). The mean number of SNPs with F_{ST} above 0.45 is 18.9%. Among the 5 high- F_{ST} genes that we studied all except one have less than 50% SNPs with a F_{ST} above 0.45 (mean 35.7%).

When considering only SNPs with heterozygosity above 0.2 this pattern remains unchanged, though the frequency of SNPs with F_{ST} above 0.45 increases slightly (from a mean of 24.1% for both classes to 39.3% – Table 9.3). Genotyping 2 random SNPs will increase the probability of detecting at least one high F_{ST} SNP above 50% for most genes. Power above 80% will require 4 to 8 SNPs per gene to detect at least one SNP with F_{ST} above 0.45.

For the five putative genes under selection, the probabilities are slightly lower than for the five high F_{ST} genes: two SNPs are rarely enough to have probability above 50% and in most cases 8 SNPs are required to have power above 80%.

				SNP F_{ST}		
Gene symbol	Function	Selection evidence	Number of SNPs	Median	Mean	Max
Candidate genes under directional selection						
LCT	Lactase persistence	Biller and Grand (1990)	61	0.119	0.260	0.756
SLC24A5	Skin pigmentation	Ginger et al. (2008)	27	0.157	0.304	0.983
KITLG	Skin pigmentation	Wehrle-Haller (2003)	124	0.041	0.170	0.775
OCA2	Skin pigmentation	Slominski et al. (2004)	542	0.135	0.186	0.871
MDM4	Apoptosis	Atwal et al. (2009)	48	0.214	0.217	0.490
Genes with high proportion of SNPs with $F_{ST} > 0.45$						
ANTXR1			196	0.127	0.232	0.820
CAB39			41	0.240	0.269	0.621
CAD			27	0.227	0.238	0.590
CLASP1			258	0.491	0.372	0.994
MAP4K3			198	0.160	0.260	0.809

Table 9.1: Characteristics of the candidate and high- F_{ST} genes studied. The table includes gene function with publications describing evidence for selection, total number of SNPs in introns and exons, and statistics (median, mean and maximum) for the F_{ST} of the SNPs.

Gene symbol	Chromos. number	Start position	Length of gene (kb)	Number of exons	Length of exons (kb)
Candidate genes under directional selection					
MDM4	1	202752133	33842	10	2027
LCT	2	136261884	49336	17	6274
KITLG	12	87410697	87672	10	5434
OCA2	15	25673615	344438	24	3140
SLC24A5	15	46200460	21421	9	1617
High- F_{ST} genes					
ANTXR1	2	69093992	133000	13	2125
CAB39	2	231285908	108123	9	3718
CAD	2	27293761	26397	44	7108
CLASP1	2	121811824	311698	39	8086
MAP4K3	2	39329925	187798	34	4113

Table 9.2: Summary information for all 10 genes used. The first five genes are known to be under directional selection. The last five have the largest proportion of SNPs (chromosome 2) with F_{ST} above 0.45. The table includes the symbol name, chromosome number, start position in relation to NCBI human genome build #36, gene length (introns and exons), number of exons and length of all exons together.

For long genes (with total length – exons and introns together – of at least 100 kb) it is possible that only certain segments of the gene have SNPs with substantially high F_{ST} 's. For example the ANTXR1 gene has 60% of all SNPs with an $F_{ST} > 0.45$ in the first 25% stretch (5' end) of the gene, but this number drops to 0% in the last 25% (3' end; Table 9.3 and Figure 9.2). Similarly, the OCA2 gene has 0% of SNPs with high- F_{ST} in the start (5') stretch.

The probability of finding a high- F_{ST} SNP on chromosome 2 while sampling 1, 2, 4 and 8 SNPs in random 150 kb length zones is respectively 3.6, 6.8, 11.8 and 22.8% (Table 9.3). Sampling known genes (including introns) has similar probabilities of 3.6, 6.8, 11.6 and 18.7% respectively. If SNPs are chosen completely at random from the entire chromosome 2 (not within a subregion) then the probability of finding a high F_{ST} SNP is respectively 6.4, 12.4, 23.3 and 41%. Again, using SNPs where heterozygosity is above 0.20 increases the probability of finding a high F_{ST} SNP. We repeated this analysis for chromosome 12 and results were qualitatively similar.

Haplotype reconstruction (Figure 9.3) suggests that using 4–8 SNPs will make F_{ST} converge to a single value per gene. This happens even when randomly choosing SNPs to construct haplotypes. The genes under selection or the genes with a high proportion of high F_{ST} SNP tend to have an haplotype F_{ST} above putative neutral genes, but this value of haplotype F_{ST} is still often below 0.45.

Finally, we verified our results by comparing overall F_{ST} for chromosome 2. While we know of no studying reporting pair-wise F_{ST} between YRI and CEU populations, our results are comparable to other reports (see e.g. Amato et al., 2009): We used 120684

Locus	All SNPs with $F_{ST} > 0.45$				$H_e > 0.2$ SNPs with $F_{ST} > 0.45$			
	% of SNPs	1 of 2	1 of 4	1 of 8	% of SNPs	1 of 2	1 of 4	1 of 8
LCT	25.0	43.8	68.4	90.0	35.1	57.9	82.3	96.9
SLC24A5	26.1	45.4	70.2	91.1	54.5	79.3	95.7	99.8
KITLG	22.3	39.7	63.6	86.7	71.4	91.8	99.3	100.0
MDM4	10.6	20.0	36.0	59.0	13.9	25.8	44.9	69.7
OCA2	10.4	19.8	35.7	58.6	13.7	25.5	44.5	69.1
OCA2 (5')	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
OCA2 (3')	14.6	27.0	46.7	71.6	22.4	39.8	63.7	86.8
CAD	29.6	50.5	75.5	94.0	40.0	64.0	87.0	98.3
CLASP1	56.1	80.8	96.3	99.9	77.0	94.7	99.7	100.0
CAB39	32.8	54.9	79.6	95.8	44.4	69.1	90.5	99.1
MAP4K3	29.7	50.5	75.5	94.0	44.3	68.9	90.3	99.1
ANTXR1	20.1	36.2	59.3	83.4	29.1	49.8	74.8	93.6
ANTXR1 (5')	60.0	84.0	97.4	99.9	77.4	94.9	99.7	100.0
ANTXR1 (3')	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean ¹	24.1	40.1	60.0	77.2	39.3	57.7	74.1	84.5
Mean ²	34.3	53.9	74.4	90.0	50.1	71.0	87.3	95.8
Random choice								
Chr 2	6.7	13.0	24.2	42.6	9.0	17.2	31.4	53.0
Chr 12	6.4	12.4	23.3	41.0	8.1	15.6	28.7	49.1
500 random chromosome samples of 150 kb length								
Chr 2	3.8	5.0	10.8	16.8	5.6	9.4	16.1	25.0
Chr 12	3.6	6.8	11.6	18.6	7.0	11.8	17.6	21.8
Random samples of entire genes (including introns)								
Chr 2	3.6	6.7	11.7	18.7	5.8	10.2	16.3	23.8
Chr 12	4.1	7.3	12.4	19.0	6.5	11.0	17.0	23.8

Table 9.3: Detection of high F_{ST} SNPs in the 10 studied genes, random choice of SNPs across a chromosome, random 150 kb samples, and random samples from all genes with more than 20 SNPs. The left half of the table includes all SNPs whereas the right side includes only SNPs with expected heterozygosity above 0.2. Each column shows the probability of finding at least one SNP with F_{ST} greater than 0.45 when sampling 1 or sets of 2, 4 or 8 SNPs. For OCA2 and ANTXR1 the initial and final 25% are studied separately. Mean¹ includes all OCA2 and ANTXR1 whereas Mean² includes only the 25% gene segment with high- F_{ST} SNPs.

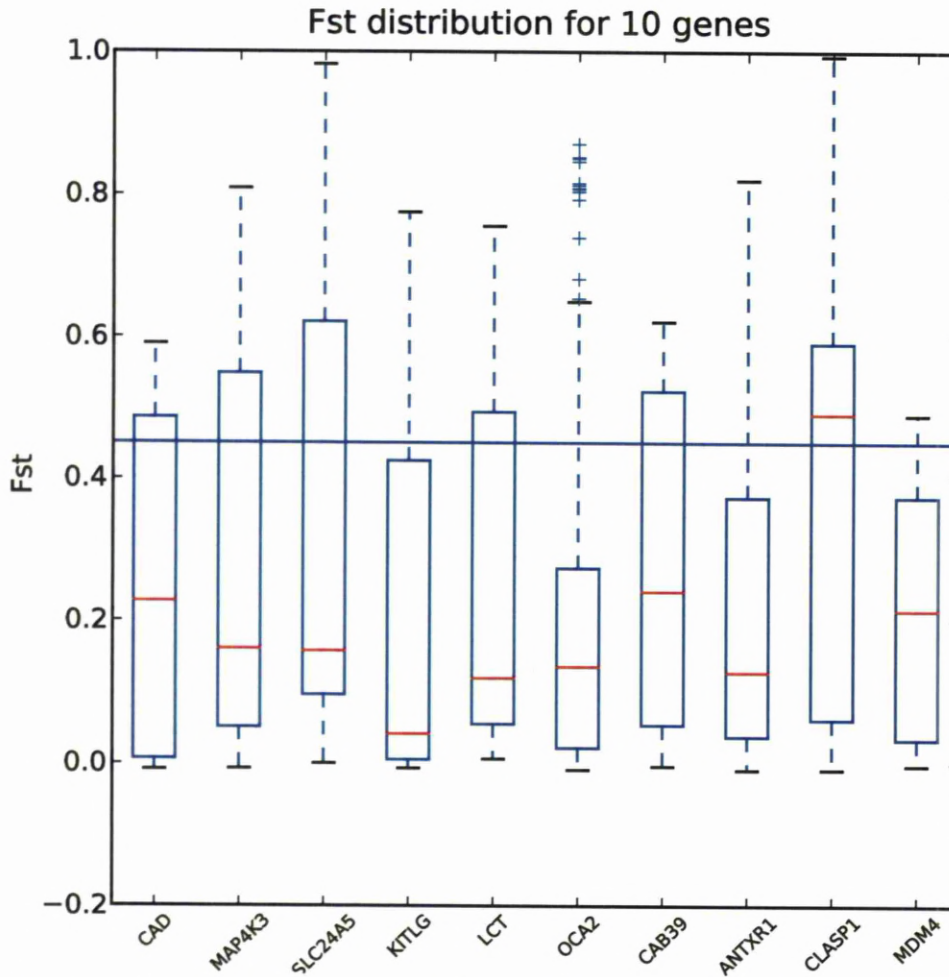


Figure 9.1: The F_{ST} distribution for all SNPs in the 5 genes under directional selection and 5 genes in chromosome 2 with the highest proportion of SNPs with F_{ST} above 0.45. Bottom and top edges of boxes are the 25% and 75% percentiles of the distribution. The horizontal bar in each box is the median (i.e. 50% percentile). The whiskers extend to the most extreme data point below 1.5 times the inter-quartile range from the box.

SNPs and obtained a median F_{ST} of 0.0712, mean 0.1181. The 0.45 point is the 96.4% percentile and 0.494 is the 97.5% percentile (for 95% confidence intervals).

9.4 Discussion

Our most important result is that genotyping only one or few SNPs per locus is often insufficient to detect high F_{ST} and directional selection associated with local adaptation

or speciation (Beaumont, 2005). We first discuss this problem of “false negatives” (i.e. failure to detect selection).

SNP sampling and false negative rates

Sampling a single random SNP is rarely enough to detect a high F_{ST} gene. Even in the case where, by definition, the gene has a high proportion of high F_{ST} SNPs, genotyping only one random SNP will give at most 56.1% probability of detecting F_{ST} above 0.45 (Table 9.3). Most high F_{ST} genes might often require at least 4 SNPs to have power near 80%. Genes that are known to be under selection will often require up to 8 SNPs in order to have power above 80%. As expected (Allendorf and Seeb, 2000) using SNPs with expected heterozygosity above 0.2 slightly improves, in most cases, the power to detect High F_{ST} genes.

Our results suggest that current sampling strategies (with only 1 or 2 SNPs per gene) with many non-model species are insufficient to detect genes with high- F_{ST} SNPs. Even with 4 SNPs sampled (genotyped), some high- F_{ST} genes likely would not be detected at 80% power.

It is not clear if F_{ST} can reliably detect selection when genotyping less than 8 (or even more) SNPs when directional selection is relatively weak. For instance research from (Atwal et al., 2009) suggests that the MDM4 gene is only under selection on CEU populations and our results suggest detection would be improbable (59.0–69.7%) with even 8 SNPs.

If genes are long (e.g. >100–200kb), having different set of SNPs targeting different parts of the gene should increase the power to detect High- F_{ST} areas. As the cases of ANTXR1 and OCA2 demonstrate that high F_{ST} gene subsegments might be concentrated in some specific part of a gene. This effect can be explained with two complementary observations: long genes will have a higher probability of spanning areas with in-between recombination hotspots and if the selection effect is ancient recombination will potentially lower F_{ST} in parts of the gene distant from the actual area under selection.

Haplotype inference

Haplotype inference tends to converge to a central value per gene as the number of SNPs used is increased (Figure 9.3), even with a random choice of SNPs. Genes with high- F_{ST} or under selection tend to converge to higher values than other genes, though not to values above the 0.45 cut for SNPs. Nonetheless it seems clear that a rank based on haplotype F_{ST} is possible: the higher the haplotype F_{ST} , the greater the probability of being a gene under selection or with high- F_{ST} . Haplotype reconstruction is, unfortunately problematic with most species: The YRI and CEU populations of the HapMap dataset are a best case scenario as pedigree information (mother, father and

offspring) is available and was used to perform haplotype reconstruction. Most other datasets (including human datasets) will not have as much information to do haplotype reconstruction and thus such an option will probably not be available in many studies.

False positive rates

While it is not possible to completely characterise the false positive rate (as the neutrality status of all segments of genome is not known) it is possible to quantify the probabilities of: (i) detecting 1, 2, 4, 8 SNPs with high- F_{ST} by random sampling; (ii) detecting high- F_{ST} SNPs in a 150 kb block and (iii) finding high- F_{ST} SNPs in all genes. These quantifications will be over-estimates of the false positive rate (as they will also include parts of the genome that are either under selection or strongly linked to areas under selection).

Finding 1, 2, 4 and 8 SNPs with High- F_{ST} by randomly sampling shows qualitatively different results when scanning 150 kb spans of chromosomes and, genotyping SNPs in known genes. Indeed, when using 8 SNPs the “false positive rate” will more than double when comparing totally random scans to inside-gene scans. This suggests that F_{ST} is not distributed randomly and that high F_{ST} SNPs tend to cluster together (as found, using a different method, in Akey et al. (2002)), this is further confirmed as a smaller window (150 kb) has less probability of finding a high SNP F_{ST} than a longer window (1.5 mb), especially when 8 SNPs are used. Both non-random genotyping strategies suggest that with 4 to 8 SNPs between 10 and 40% of all observations would report back a high- F_{ST} location. Discounting real observations of selection, the false positive rate will probably be still above 5% with 4 SNPs and probably above 15% with 8 SNPs. Even so, with an average coverage of 8 SNPs or less per gene, and assuming that the highest F_{ST} SNP found as a representative for the gene, F_{ST} might be a reasonable strategy in the face of the low information available to detect genes under directional selection.

Gene targeted approaches are thus beneficial to detect selection because the false positive rate is lower in genes than in random SNPs or even chromosome segments and selection is often directed at genes. Furthermore, gene targeted SNP discovery is increasingly feasible in non-model species (Cosart et al., In press).

SNP genotyping and discovery

This research questions the usefulness of increasingly available high density SNP chips (micro-arrays) for detecting selection when the SNPs are evenly distributed across the genome. For instance, the cow 54k SNP chip design of Matukumalli et al. (2009) with a minimum gap of 22.5 kb between SNPs would often be insufficient because <4 SNPs would occur in many genes (e.g. in 4 out of the 10 studied genes we studied here). SNP chips with shorter gaps between SNPs (e.g. <10 kb between SNPs) could be developed

to increase the power to detect selection. For chip design, SNPs in genes could also be favoured to allow a gene-targeted approach to increase the probability of detecting selection.

Exon capture provides an ideal SNP discovery and genotyping technology for detecting selection using either a candidate gene approach or a genome wide scan approach using gene-targeted SNPs (Hodges et al., 2007). For example, exon capture micro-arrays can be designed to allow for next generation sequencing of a few exons from each of tens to thousands of genes (Cosart et al., In press). This would allow genotyping of 4 to 8 SNPs per gene if at least approximately 2 kb were sequenced per gene (assuming one or more SNPs per 500 bp as is common in natural populations (Morin et al., 2004; Hohenlohe et al., 2010)). Exon capture also allows targeted sequencing of 3' and 5' exons (Cosart et al., In press), which helps avoid missing selection signatures occurring at only one end of a gene as in ANTXR1 (Figure 9.2 c).

Limitations

Our analysis is subject to limitations. First, we only considered pair-wise F_{ST} as we only include two populations in our study.

Secondly, calculating the false positive rate across entire chromosomes or gene sets is not trivial as the selection status of most genes is not known. Therefore we can only provide the probability of finding high- F_{ST} SNPs using different sampling strategies and from there conduct an informed assessment of the possible false positive rate of different genotyping strategies.

Our work is highly dependent on the quality of both the Known Gene and the HapMap databases. The HapMap project dataset is known to suffer from ascertainment bias (Clark et al., 2005) and, for association tests between SNPs and complex disorders the power should be slightly eroded. This problem can be addressed with the complete sequencing of genomes (e.g., the 1000 genomes project <http://www.1000genomes.org/>). Most importantly, the Know Genes database included in some cases and for the same gene different alternatives for gene size and exon number and position. Manual inspection suggested that differences were minor in most cases and that the quality of the curation for this dataset is, and will be for the foreseeable future much better than similar information for non-model species (if it is available at all).

Here we focused on two distinct problems: (i) the ability to detect high- F_{ST} genes when they exist and (ii) detection of high- F_{ST} SNPs in genes that are under selection. For point (i), it is clear that we used only a small subset of genes (5), but such a limited subset is enough to address the fundamental question that even for genes with a high percentage of high- F_{ST} SNPs, detection clearly requires more SNPs than is commonly used for non-model species and limited genome wide scans and candidate gene approaches. For point (ii), it is less clear what has been the actual mode(s), timing, and

strength of selection on the genes studied, therefore it is not completely clear what the impact of selection on F_{ST} would be other than there was probably directional selection occurring.

Some limitations are self-imposed. The purpose of this study is to use an empirical dataset for which there is both many markers and knowledge of demographic history and selection status of many genes in order to develop understanding and recommendations applicable to species where much less information is available. Therefore we sometimes self-imposed limitations like assuming that nothing is known e.g. about gene locations across the genome or the function of an SNP (intron or exon based, synonymous or non-synonymous).

9.5 Conclusions and recommendations

Due to the high number of SNPs required to have reasonable power and the relatively high false positive rate, it is difficult to provide clear quantitative guidelines as to genotyping SNPs to detect molecular signatures of adaptive differentiation between populations. Nonetheless a few helpful guidelines can be made:

1. Power to detect selection is usually low with less than 4 SNPs per gene, because variance in F_{ST} among SNPs is generally high within a gene. Therefore we recommend the analysis of at least 4 SNPs per gene.
2. Targeting genes or gene rich areas is a more reasonable strategy than a completely random genomic scan using only hundreds or a few thousand SNPs to avoid failure to detect high F_{ST} SNPs.
3. Each individual researcher will require balancing between failing to detect selection signals and detection of false positive signals. Depending on the research project, different false positive and negative rates might be acceptable.
4. All the information available should be used. If known, gene position, gene function and mutations relevant to protein conformation should be used.
5. For long genes, genotyping sets of SNPs at the ends may increase the probability of finding high- F_{ST} SNPs because recombination and drift could prevent selection from driving up F_{ST} at both distant ends of genes.
6. Design of high density SNP chips or assays should include several SNPs per gene. Researchers using existing SNP assays should verify the distance between SNPs (e.g. at least <20 kb in gene-rich areas) when using a candidate gene approach or genome-wide scan.

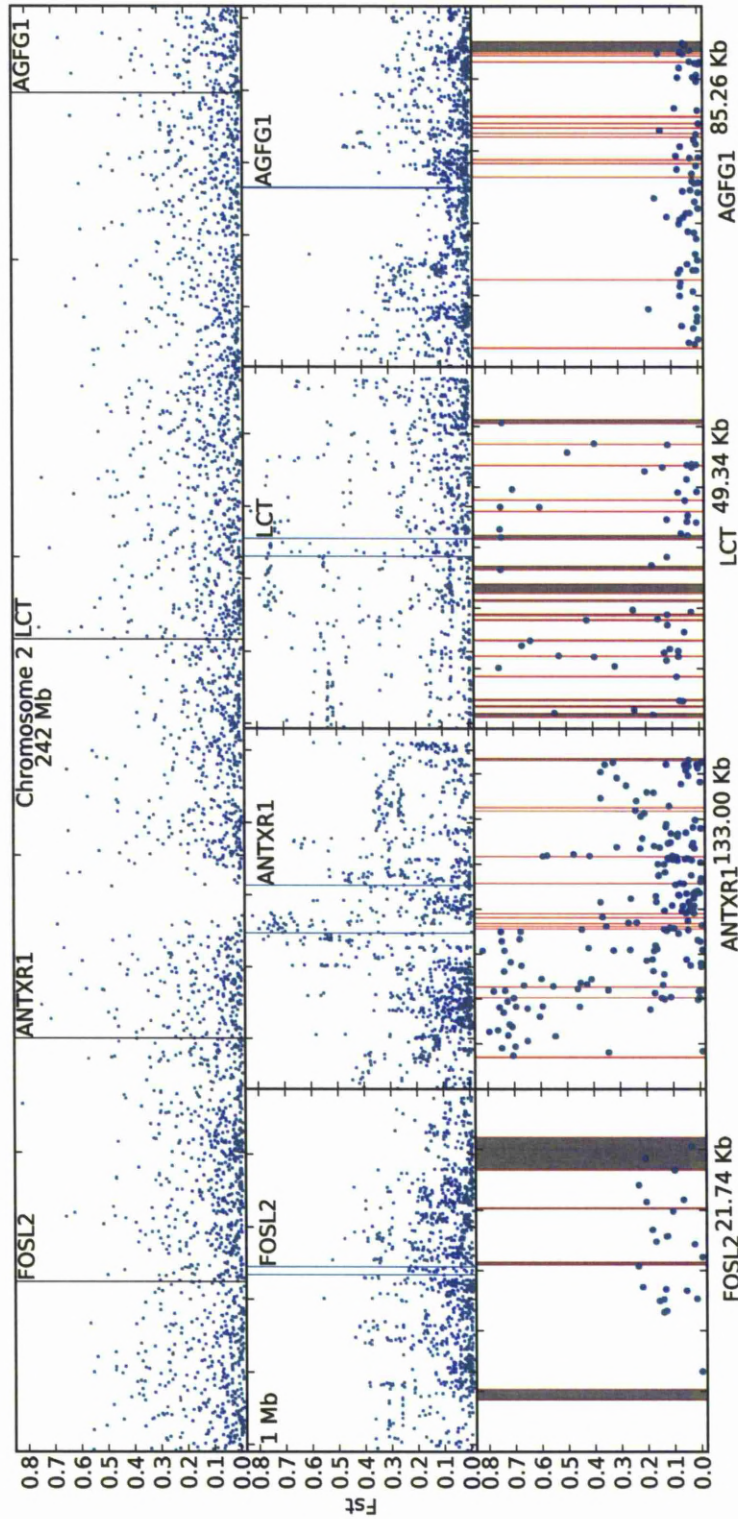


Figure 9.2: F_{ST} for 1% of SNPs along chromosome 2 and zooming in on four genes sampled from Yorubans in Africa and Utahans representing North Western Europeans. The first row shows F_{ST} along the whole chromosome (242 Mb). The second row covers a area of 1Mb centred around the four genes. The last row shows the F_{ST} for all SNPs in each gene, the gray boxes represent exons (for example FOSL2 has four exons). FOSL2 (21.74 kb) and AGFG1 (85.26 kb) represent “neutral” genes, ANTXR1 (133.0 kb) represents a gene with many high F_{ST} SNPs and LCT (49.34 kb) is a candidate gene known to be under directional selection.

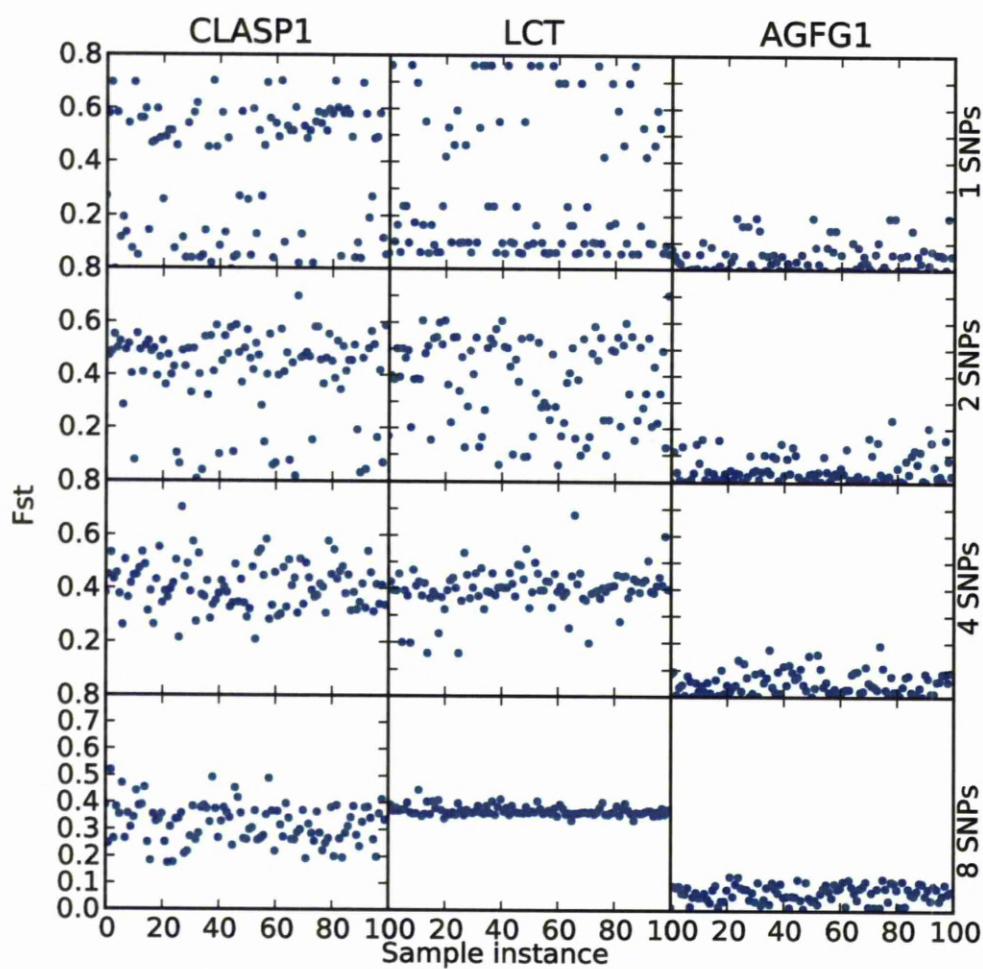


Figure 9.3: F_{ST} for random haplotype reconstructions using increasing amounts of SNPs. Row 1 shows the sampling of “haplotypes” with a single SNP. Row 2, 3 and 4 with 2, 4 and 8 SNPs respectively. The X-axis represents a haplotype reconstruction using randomly selected SNPs.

Ten

Assessing selection for drug resistant malaria: Can temporal F_{ST} help?

Tiago Antao, Gordon Luikart and Ian M. Hastings

Abstract

Plasmodium falciparum malaria is treated by antimalarial drugs which select for drug resistant *Plasmodium* parasites. Detecting which genes encode resistance can provide fundamental insights to help manage and monitor the spread and the burden of drug resistant *P. falciparum*. F_{ST} outlier tests are commonly used to find genes under selection and applied to the case of *P. falciparum*: where samples from the same population are compared over time to test for an increase in temporal F_{ST} (at resistance genes) following antimalarial treatment of a host population. We use computational simulations to evaluate the performance of temporal F_{ST} to detect genes under selection in *Plasmodium*. Our results suggest temporal F_{ST} is more powerful in low transmission areas as long as the time between samples is low (<20 *P. falciparum* generations). Temporal F_{ST} cannot reliably detect genes under selection with small changes in prevalence ($<10\%$) between samples. For example, temporal F_{ST} might not be the best strategy to find Artemisinin related resistance genes as the current prevalence of Artemisinin resistance is still low. Temporal F_{ST} is applicable in epidemiological scenarios where a large change in frequency of resistance is expected to occur, e.g., when a drug is removed or applied at high intensity. Empirical studies would also benefit from increasing sample sizes of individuals and loci, which is now feasible with new genomic technologies.

Keywords: malaria, selection, F_{ST} , drug resistance, population genomics, infectious diseases, genetic monitoring

10.1 Introduction

Malaria is a major public health concern for the one third of the global human population estimated to be exposed to the most virulent parasite species, *Plasmodium falciparum*.

Estimated numbers of clinical episodes of *P. falciparum* malaria range from 300 to 660 million per annum (Snow et al., 2005). There is no effective vaccine for this species. Infection is controlled by insecticides targeted at vector mosquito species, and treatment by antimalarial drugs. As might be expected, drug-resistance rapidly evolved and spread (Olliaro, 2005).

Determining which genes are involved in drug resistance is thus fundamental both for control and elimination efforts and has been successfully done for older drugs like Chloroquine or SP (Olliaro, 2001) though the mechanisms of resistance in the newest class, the Artemisinins still eludes us (Rogers et al., 2009; Dondorp et al., 2009). Even for older therapies, there are still doubts that all genes involved in drug resistance have been identified. Genes involved in drug resistance have asymmetrical importance, for instance, the *crt* gene is more important than the *mdr* gene in conferring resistance to Chloroquine and the same observation was made for *dhfr* and *dhps* with regards to SP resistance.

There are several strategies to detect genes under selection, among them multilocus comparisons. The Lewontin-Krakauer test (and more recent versions) utilises the variance in F_{ST} estimates from multiple independent loci to determine if non-neutral markers are present: Loci that undergo different directional selection in some populations are expected to show larger allele frequency differences among populations. F_{ST} studies and subsequent selection tests are commonly done over many species and *P. falciparum* is not an exception (e.g. Escalante et al., 2001; Iwagami et al., 2009; Anderson, 2004).

Multilocus selection tests are normally conducted comparing different populations, but they can be applied to one population, sampled over time (for examples with *P. falciparum* (see e.g. Abdel-Muhsin et al., 2003; Chenet et al., 2008; Gatei et al., 2010). These tests can be used to detect the impact of many artificial interventions ranging from the deployment and removal of new drugs (Abdel-Muhsin et al., 2003) to the impact of usage of insecticide treated bednets (Gatei et al., 2010).

Here we conduct computational simulations of the spread of drug resistant malaria to understand how temporal F_{ST} signals are influenced at drug resistant loci. Some of the questions we address are: “What increase in prevalence is required before directional selection can be detected?”, “Is the F_{ST} signal similar at loci with asymmetrical (or unequal) contribution to resistance (e.g. comparing *dhfr* with *dhps*, the loci involved in SP resistance)?”, “When can selection tests based on F_{ST} be used to find the genes responsible to Artemisinin resistance?”, “Are F_{ST} tests more informative in areas of low- or high-transmission?”, “Can F_{ST} be informative with seasonal population fluctuations (like the ones imposed in areas with wet and dry seasons)?”, “Are commonly used sampling strategies enough to provide a reliable estimate of F_{ST} ?”

The World Health Organisation (WHO) recommends a change in antimalarial treatment if the total treatment failure proportion is equal or above 10% (World Health

Organization, 2006). This suggests that any simulations of *P. falciparum* drug resistance should concentrate on frequencies of resistance that span a range from zero up to 10% failure rate. We will assume that the treatment failure rate will be equal to the prevalence of resistance (i.e. the proportion of humans that harbour resistant infections). The prevalence of resistance is defined as the number of individuals that harbour resistant infections. An important parameter in *P. falciparum* biology is thus the Multiplicity Of Infection (MOI), i.e. the number of different infections that a human host carries. It should be clear that prevalence resistance should not be confused with the frequency of resistance as individuals might have more than one simultaneous infection, so for any given frequency, prevalence increases with MOI.

As *P. falciparum* parasites only reproduce sexually inside the mosquito vector, mating opportunities are limited to the parasites ingested in the last blood meal (mosquitoes feed approximately every three days so mating between parasites in blood meals obtained in separate bites is assumed to be impossible) thus the MOI is one of the fundamental factors influencing inbreeding. Selfing (mating with a genetically identical parasite) is possible. The MOI is a proxy for transmission intensity: due to repeated sequential infection typical in high transmission areas, infected individuals will have a higher MOI than in low transmission areas. Higher transmission areas will thus have lower parasite inbreeding (e.g. selfing) levels.

10.2 Methods

We performed two different kinds of simulations: One approach used a simulator of malaria epidemiology and population genetics to study genes under selection due to drug pressure. This simulation exercise is intended to estimate the plausible values of temporal F_{ST} with varying prevalence and frequency of resistance.

A second set of simulations used a standard forward-time population genetics simulator, in order to establish the expected null distribution of temporal F_{ST} with neutral markers. This simulation exercise is intended to answer a different set of questions, namely what is the impact of sampling strategies on the estimation of neutral temporal F_{ST} .

Simulation of genes involved in drug resistance

We simulated data using ogaraK (Antao and Hastings, 2011a), a population genetics simulator of the spread of drug resistant *P. falciparum* using an infinite population size model. OgaraK is able to simulate the non-standard biological features on *P. falciparum* biology, namely: (i) a genome that is haploid (with asexual reproduction) but with a brief diploid (sexual) phase inside the mosquito vector; (ii) mating options inside the mosquito are limited and dependent on MOI (see below) and (iii) infections,

not individual parasites are simulated due to the very large number of individuals (up to 10^{12} in a single human host). A formal description of the model and application is available in Hastings (2006); Antao and Hastings (2011a,b). Here we present the fundamental concepts and parameters related to this study.

Parasite resistance is modeled using 2 loci that encode drug resistance (for a more general model with different number of loci see Antao and Hastings (2011b)). Each locus has 2 alleles (sensitive and resistant). Resistance to drug treatment might require both loci being resistant (full epistasis); or only a single locus being resistant (duplicate gene function) or one specific locus (of the 2) being necessary and sufficient to confer resistance (asymmetrical epistasis). Asymmetrical epistasis is based on Chloroquine and SP resistance where two loci were identified in resistance but one loci is more important than the other (*crt* is more important than *mdr* in Chloroquine resistance and *dhfr* is more important than *dhps* in SP resistance). Resistance alleles incur a fitness penalty in the absence of the drug and all mutations are assumed to have the same fitness penalty. Genotypes with multiple mutations suffer a multiplicative fitness penalty.

The mode of resistance required depends on the human: (i) in untreated humans all parasites survive drug treatment; (ii) in humans with no immunity or with poor treatment compliance a weak mode of resistance (i.e. asymmetry or DGF) is enough for a infection to survive treatment and (iii) in semi-immune humans with complete treatment compliance only the strong mode (full epistasis) is enough to confer resistance.

Assuming the resistance modes above we simulated three different scenarios:

Simple Full Epistasis: All parasites in untreated individuals survive. Parasites in treated humans require both alleles to be resistant.

Simple DGF: All parasites in untreated individuals survive. Parasites in treated humans require only one allele to be resistant.

Mixed mode: All parasites in untreated individuals survive. Treated humans are split in two groups of equal size: The first group, representing individuals with no immunity and/or insufficient drug doses will allow parasites to survive with asymmetrical epistasis (i.e. the most important resistant locus is necessary and sufficient to confer resistance). The second group, representing individuals with semi-immunity and/or sufficient drug treatment will allow only parasites with both resistant alleles (full epistasis), to survive.

In simple modes, the parasite genotype is enough to determine the survival to drug treatment, whereas in mixed mode, the environment (human immunity status and/or treatment compliance) is also a factor.

For all simulation scenarios we varied the fitness penalty per resistant mutation and the amount of drug usage (defined as the percentage of infected people that is

treated) between 0 and 20% in increments of 2%. We simulated 5 different MOIs: three simple models with MOI fixed at 1 and 2 (low transmission) and 4 (high transmission) and two more realistic scenarios, one modeling low transmission where 50% of human hosts had a single infection and the other half had 2 infections and another modeling high transmission where MOI followed a Poisson distribution with a conditional mean of 2.3 truncated at a maximum MOI of 7 Hastings (2006). The simpler MOI models qualitatively capture the results of the more complex ones therefore we will present all results with fixed MOIs. Initial frequency of resistance was set at 0.1% for each resistance allele and no linkage disequilibrium among resistance loci.

Linkage disequilibrium (LD), the non-random association of resistance alleles at different loci, has been shown to be a critical factor influencing the rate of spread of resistance (Dye and Williams, 1997). Here we use r as a measure of LD as the signal of r gives important information about the type of association encountered: a positive r means that resistance alleles at different loci are more likely to be associated than expected by random association, whereas a negative value suggests that resistance alleles are less associated than expected.

While we have complete information about the parasite population, in order to present a more realistic estimate of F_{ST} , we sample 100 haploid individuals and use Cockerham and Wier's θ (Weir and Cockerham, 1984) implemented in Genepop (Rousset, 2008). In order to do the many thousands of evaluations needed we used the Bio.PopGen module of Biopython (Cock et al., 2009) to automatise the F_{ST} estimator calculation. We assume that we can trivially infer haplotypes as such information is available from simulated data, but we notice that, in real settings haplotype inference is not trivial when $MOI > 1$.

We considered that resistance existed at the onset of the simulation (i.e. this study is not concerned with *de novo* emergence of resistance) with a starting frequency of 0.1% and ran simulations for 100 generations (approximately 20 years for *P. falciparum*).

Simulation of neutral markers

We conducted simulations using the forward-time, individual based simulator simuPOP (Peng and Kimmel, 2005). Simulations were performed using a Wright-Fisher model with separate sexes, random mating (average sex ratio of 1) and discrete, non-overlapping generations. We simulated constant size populations with an N_e of 1,000, 5,000 and 20,000, in line with estimations of N_e for low-, medium- and high-transmission areas respectively (Anderson et al., 2000a). Each demographic scenario was replicated 1,000 times. Simulations had a burn-in phase of at least 10 generations in order to approximate mean observed heterozygosity with realistic values (below 0.8) and lasted 100 generation. Longer burn-in periods were also tested, but results were qualitatively unchanged. The genome simulated included 100 neutral, independent microsatellite loci

initialised with a Dirichelet distribution (10 initial alleles per locus exhibiting a mean of 8 after burn-in) and no mutation rate. As with the simulations of loci involved in drug resistance, simulation data was saved in the Genepop (Rousset, 2008) format and automatically processed using Biopython (Cock et al., 2009).

10.3 Results

Simulation of genes involved in drug resistance

The frequency of resistance alleles is not clearly related with the frequency of resistance (phenotype) and is dependent on MOI and epistasis mode (similar results were obtained in Antao and Hastings (2011b)). While for full epistasis there is a direct relationship between allele frequency and resistance phenotype (i.e., the increase in frequency of the haplotype with all resistance alleles is parallel to the increase in frequency of individual resistance alleles – Figure 10.1), the pattern is different for the other two scenarios: in mixed mode, the most important locus, as expected, is fundamental in determining the increase in resistance phenotypes. Indeed, any temporal F_{ST} signal for the least important locus is delayed when compared with the spread of resistance phenotypes. With DGF the direct relationship between resistant allele frequency and resistance phenotype is even weaker: it is possible that resistant alleles stabilise at intermediate frequencies whereas the resistance phenotype approaches fixation through an increase in absolute value of linkage disequilibrium (in the case of DGF and using r as a measure, r is increasingly negative).

To understand the epidemiologically relevant allele frequencies it is necessary to establish a relationship between prevalence and frequency due to WHO policy recommendations being based on prevalence, not frequency. The prevalence is given by:

$$1 - (1 - R)^i \quad (10.1)$$

Where R is the frequency of resistance and i is the MOI. This simplified formula is easier to compute with Full Epistasis and DGF (not mixed mode) and using a fixed MOI.

MOI has a considerable impact on prevalence for very low frequency of resistance: For very low frequencies, the prevalence is approximately MOI times larger than the frequency (Figure 10.2). The relative impact of MOI on prevalence is highly reduced as frequency increases.

As per Antao and Hastings (2011b) $R = F_{1,1}$ for full epistasis, where $F_{0,0}$ is the frequency of the haplotype with no resistant alleles, $F_{1,0}$ the frequency of the haplotype where the first loci is resistant and the second sensitive, $F_{0,1}$ the frequency of the haplotype where the first loci is sensitive and the second resistant and $F_{1,1}$ is the frequency

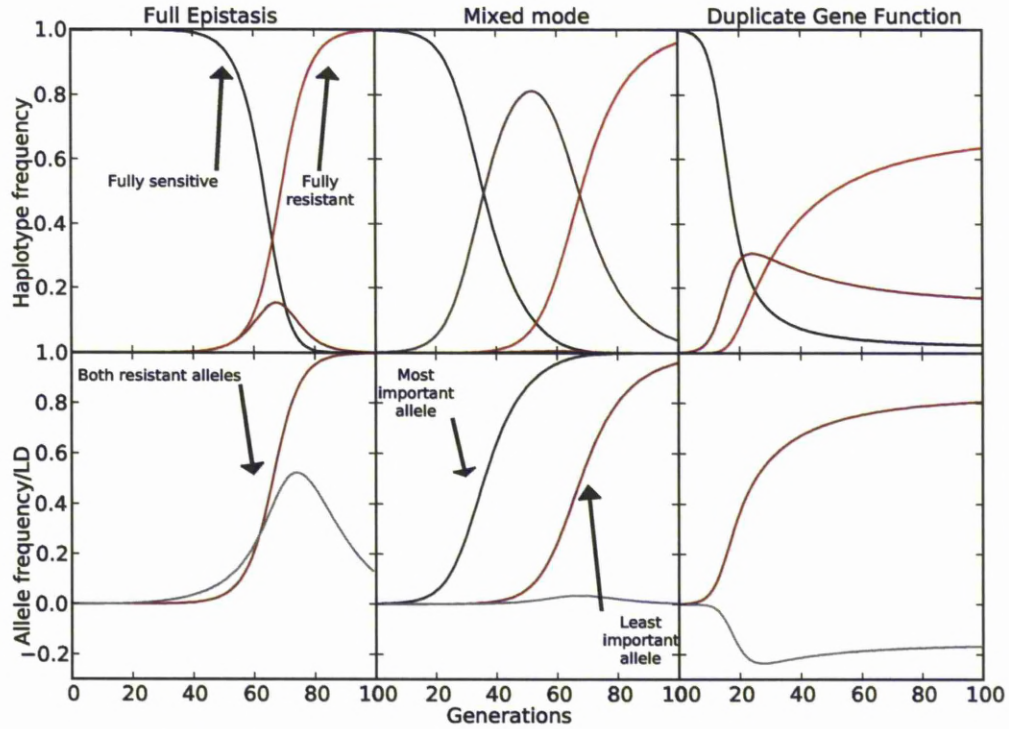


Figure 10.1: Summary statistics for simulations using the three epistasis modes. Each column documents a different epistasis mode over time (generations since the start of the simulation). The first row shows haplotype frequency (Lighter tones of grey indicate an increasing number of resistance alleles in the haplotype). The second row shows allele frequency, in the case of symmetrical modes (full epistasis and duplicate gene function – DGF) the frequency for both alleles is equal, but in mixed mode the frequency of the most important gene increases before the frequency of the least important gene. The third row also shows LD (r – grey line) between loci involved in drug resistance. The drug treatment rate is 20% and the fitness penalty 5%.

of the haplotype having all resistant alleles. $F_{1,1}$, is thus, assuming full epistasis, the only resistant haplotype (the two resistant alleles are required).

For DGF, the frequency of resistance is $F_{1,1} + F_{0,1} + F_{1,0}$, as it is enough to have only one resistant allele to survive treatment. For mixed mode, the calculation is slightly more complex and an intermediate value is expected as in some cases (treated non-immune humans and/or with poor drug compliance) a single mutation $F_{1,0}$ (but not $F_{0,1}$) is enough to confer resistance, whereas in others (semi-immune humans and/or proper drug compliance) both mutations are required, thus assuming equal proportions of individuals in both groups, $R = \frac{F_{1,0} + F_{1,1}}{2} + \frac{F_{1,1}}{2}$. It is now trivial, to determine the relationship between the frequency of resistance of a single locus and prevalence (knowing MOI and LD).

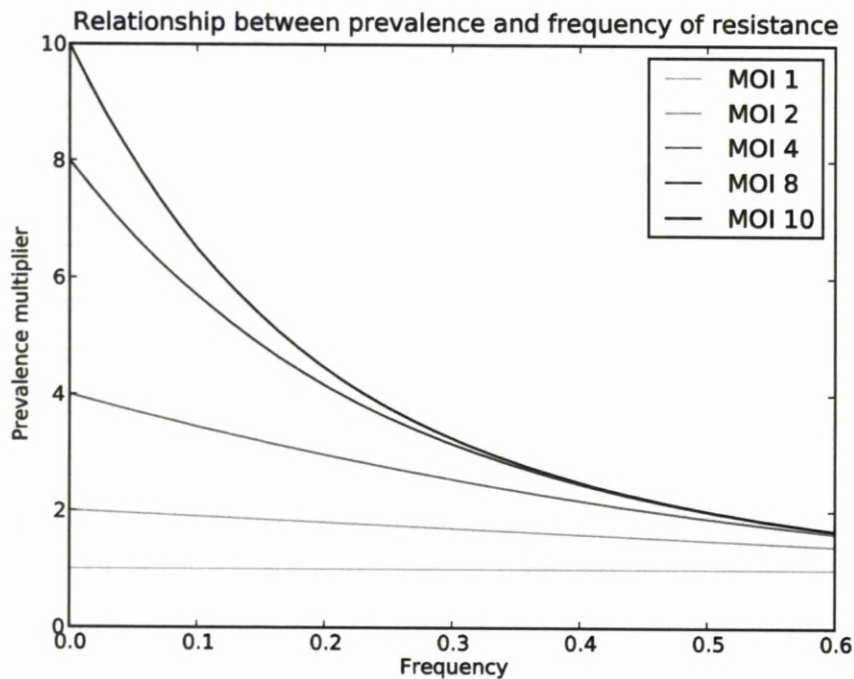


Figure 10.2: The ratio between frequency of resistance and its prevalence as a function of the multiplicity of infection (MOI). For the same frequency, higher MOI entails higher frequency. Most importantly, the ratio between prevalence and frequency clearly shows that MOI is a much more important factor, in terms of prevalence, at lower frequencies.

With the above calculation we can now study pairwise F_{ST} between time samples and know the difference in prevalence and frequency in those samples. Figure 10.3 shows pairwise F_{ST} between samples where prevalence varies between 0 and 20%. Full epistasis shows an higher F_{ST} than DGF for the same MOI and difference in prevalence: a smaller frequency of resistance alleles is required with DGF to confer resistance (as long as $LD - r$ is negative) because either gene is enough to confer resistance, therefore for the same prevalence F_{ST} is lower in DGF due to lower frequency.

The second factor influencing temporal F_{ST} is the MOI: an higher MOI will imply lower F_{ST} . This happens because with lower MOI a higher frequency is needed to attain the same prevalence (especially at low prevalences – Figure 10.2).

When sampling the generation with resistance near 0% and for differences of prevalence of up to 10%, F_{ST} is only slightly above 0.1 with Full Epistasis and a MOI of 1, and only near 10% prevalence. For all other modes F_{ST} is between 0.0 and 0.1. Even at 20% difference in prevalence many simulation results fall below 0.1.

If the initial generation for temporal F_{ST} has higher prevalence than the signal is further reduced. The reduction in F_{ST} is not just proportional to the prevalence of the first sample but slightly above (Figure 10.3). For example the temporal F_{ST} between prevalence 0 and 10% is larger than between 1 and 11% which is much larger than between 5 and 15%.

Simulation of neutral markers

For multilocus selection tests based on F_{ST} , the distribution of neutral loci is fundamental to make inferences about locus potentially under selection. Directional selection is expected to increase the F_{ST} of selected genes, therefore we will study the upper confidence interval of the distribution of neutral loci, because most tests assume precisely that directional selection loci is indicated by an F_{ST} above most (typically 95 or 99%) neutral loci.

Neutral temporal F_{ST} will increase over time due to drift. Figure 10.4 shows the behaviour of the upper confidence interval (95 and 99%), the mean and median values have the same behaviour (results not shown). The increase in F_{ST} is influenced by the population size: the larger the population size, the lower the increase in F_{ST} . This relationship is mostly due to drift: smaller populations (i.e. lower transmission areas) will have higher drift, thus increasing the difference between samples over time.

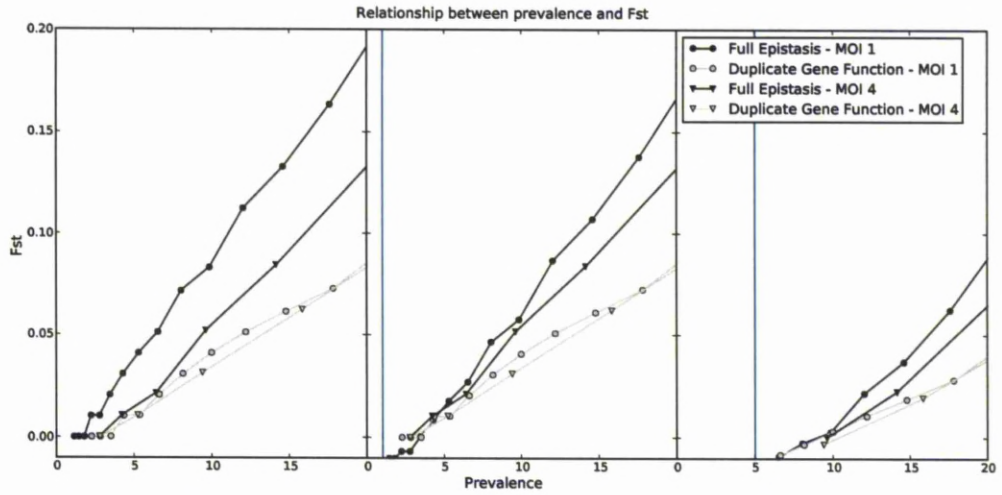


Figure 10.3: The relationship between prevalence, epistasis mode, multiplicity of infection (MOI) and temporal F_{ST} . The figure shows the pair-wise F_{ST} where the first sample is, from left to right: generation 0 (approximately 0% prevalence as frequency is at the initial 0.1%), 1% prevalence and 5% prevalence. Two MOIs (1 and 4) and two epistasis modes (full epistasis and duplicate gene function) are considered. The drug treatment rate is 20% and the fitness penalty 5%.

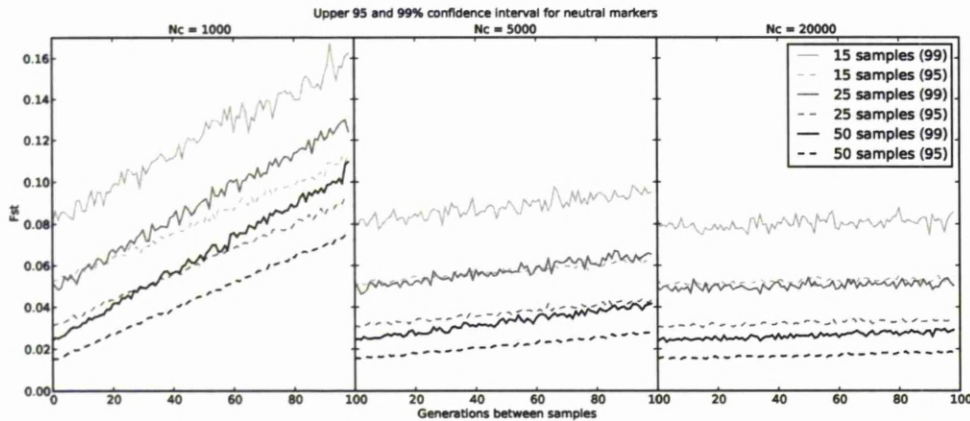


Figure 10.4: The upper confidence interval of temporal F_{ST} for neutral markers. Three different simulations with constant N_c are presented (1000, 5000 and 20000) representing values for different malaria transmission intensities. Sample sizes varies from 15 to 50 individuals (equivalent to 30 to 100 haploid *P. falciparum* clones) and is representative of several empirical F_{ST} studies. 95 and 99 confidence intervals are presented. Only the upper confidence interval is presented as this is the relevant interval for directional selection (which is expected to raise F_{ST}).

Sample size is a fundamental parameter influencing the confidence interval around estimated F_{ST} . 15 samples (30 haploid samples) will have roughly four times higher F_{ST} than 50 samples (100 haploid samples) if comparing samples that are close temporally (i.e. below 10 generations of separation). This effect will reduce over time, but even with an interval of 100 generations, the estimation using a smaller sample will be 60% higher for a population size of 1,000 and, due to less drift, 3 times more for a population size of 20,000.

10.4 Discussion

F_{ST} estimation cannot reliably identify loci involved in drug resistance under selection assuming epidemiological scenarios with gradual change. The WHO recommends that first-line therapies should be removed after 10% clinical failure rates. Assuming that clinical failure is well approximated by prevalence of resistance, sampling two time points below 10% of prevalence will yield, in most cases temporal F_{ST} below 0.1 consistent with locus neutrality. Indeed a signal of apparent balancing selection could even appear if either the generation interval is too short and/or the sample size is insufficient (while we did not simulate balancing selection, it is quite clear that simulated F_{ST} of genes involved in drug resistance can be very low). This applies to all situations where very gradual change in frequency of resistance occurs, including control interventions that might strongly reduce parasite numbers but do not target specifically resistance loci.

“Gradual” here is a relative definition: even strong selective sweeps are termed “gradual” when compared with changes imposed by seasonality (see below).

More specifically, as the initial spread of drug resistance is predicted to be very slow on an arithmetic scale, temporal F_{ST} will not be a good statistic to detect any loci under selection in the initial phases after emergence of drug resistance. This result is particularly important in the current context where resistance to artesunate is starting to spread: while the genes that are involved in artesunate resistance are not known, temporal genome-wide F_{ST} scans will probably not provide a useful way to find such genes, at least at an early stage.

F_{ST} can provide a reliable signal if the frequency change in resistance alleles is large (above 10%) and fast (e.g. within a year). Such patterns can be observed, e.g., in seasonality studies (Abdel-Muhsin et al., 2003; Babiker et al., 2005) for genes involved in drug resistance. This suggests that seasonality induced changes in the frequency in drug resistance (with the inherent changes in transmission intensity and, especially, drug treatment rate) can serve as a good scenario to better understand the dynamics of drug resistance and to better understand, if drug treatment is removed, if alleles involved in drug resistance do entail a fitness penalty for parasites in absence of drug pressure.

Sample size highly influences estimated F_{ST} . Common sample sizes used with *P. falciparum* (see e.g. Abdel-Muhsin et al., 2003; Chenet et al., 2008; Gatei et al., 2010) vary between 30 and 70 haploid (i.e. equivalent to 15 and 35 diploid) samples per unit of time. The time span in empirical studies varies from a couple of generations (studies involving seasonality) up to 40 generations (roughly 8 years assuming 5 parasite generations per year). Researchers should be particularly careful in interpreting F_{ST} with small samples (below 60 haploid individuals) and large time-spans: Our results suggest (in consonance with previous research (Kalinowski and Waples, 2002)) that low sample sizes will severely increase the variance with (neutral) loci that have low F_{ST} , furthermore drift will increase the size of the confidence interval over time. Even with studies spanning small time spans (a single generation), small sample sizes (which have been used in recent empirical studies) can account for a temporal F_{ST} of 0.08, whereas a larger sample size would decrease the confidence interval to 0.02. Loci involved in drug resistance would thus be inside the confidence interval for neutrality and could not be distinguished from neutral loci.

The estimates of temporal F_{ST} are dependent on the frequency of resistance of the first time sample: If the first sample is taken when prevalence/frequency is minimal, it maximises the value of F_{ST} . Most samples are normally taken after frequency has attained observable values, therefore the estimated temporal F_{ST} will normally be lower. This does not depend only on the difference of prevalence between samples but on the initial value itself: the lower the frequency, the higher the temporal F_{ST} . At an extreme if the two samples are taken during a stabilisation period, it is possible that temporal F_{ST} will be interpreted as balancing selection. During stabilisation, balancing selection

is indeed occurring – so the result is not in error – but the F_{ST} signal obtained will be useless to find genes involved in drug resistance.

Temporal F_{ST} for drug resistance genes will be slightly higher in low-transmission settings, therefore it should be easier to differentiate from neutral loci. But, as drift is expected to be higher in low-transmission (smaller population) areas, if the time span between samples is large, then the F_{ST} of neutral loci will increase at a faster pace than in high-transmission areas.

We considered the optimal situation where we always know haplotype frequencies. In particular a problem with particular importance in high-transmission areas comes from not being easy to establish which combinations of alleles, genotyped from a human with multiple infections, correspond to each infection. For example, a patient might be infected with two infections, one resistant at only one gene and another resistant at the other gene. In this case it is not easy to discriminate the situation above from one where a patient has one infection that is totally sensitive (no resistant alleles) and another totally resistant. This problem increases in complexity with MOI (it is not a problem when MOI is 1, i.e., there is a single infection). Thus, with higher MOIs it might not be always possible to infer haplotype frequencies from allele frequencies.

Mathematical models of malaria epidemiology and drug resistance consistently predict that once drug resistance arises, it spreads rapidly to 100% (Curtis and Otoo (1986), Cross and Singer (1991), Dye and Williams (1997), Hastings (1997), Hastings and D'Alessandro (2000)). This expectation, did not found support in field evidence which has shown, at least in some cases, that resistance may stabilise at levels well below 100% (Plowe et al., 2004; Babiker et al., 2005; Ursing et al., 2007). Hastings (2006) proposed that intense competition between separate *P. falciparum* clones co-infecting the same human can explain this observation. Antao and Hastings (2011b) suggested that epistasis between genes involved drug resistance would also be necessary to explain stabilisation at intermediate frequencies, since the results in Hastings (2006) would only happen in a very small part of the parameter space (i.e. they would be a mathematical rarity). Our research here suggests that the explanation in Antao and Hastings (2011b) might also be insufficient, because while Antao and Hastings (2011b) predicts stabilisation at intermediate frequencies of resistance alleles in many scenarios, the resistance phenotype would still approach fixation. This effect can be observed on Figure 10.1 (DGF) where haplotypes conferring resistance (top right panel) approach fixation whereas allele frequency is stabilising at intermediate frequencies. Our results suggest that existing explanations for the widely observed effect of stabilisation of resistance at intermediate frequencies should be revisited.

10.5 Conclusion

In the case of malaria epidemiology, temporal F_{ST} -outlier approaches to detect selection can be an effective tool only in very limited circumstances. It is not clear, for instance, that it will be useful in search for genes involved in artesunate resistance, especially during the initial phase of resistance spread. On the other hand, we see a strong signal of selection in some *P. falciparum* studies with varying transmission intensities (and parallel variation in drug usage rates) over the year induced by wet-dry season variability. This suggests that mutations related to drug resistance might indeed pay a considerable fitness cost or, at least, that scenarios involving seasonality such be further researched as the epidemiological and evolutionary forces present might shed considerable light regarding the spread of drug resistance.

In cases where temporal F_{ST} -outlier detection might be applicable, researchers should be careful with the sample sized used: at least 50 haploid samples should be used. In low transmission cases, a long time span between samples might increase the size of the confidence interval for neutral loci, whereas in high transmission scenarios haplotype inference will be non-trivial. If such guidelines are observed, F_{ST} -outlier approaches might provide important information regarding genes involved in *P. falciparum* drug resistance.

Part III

Estimating effective population size and assessing the success of control and elimination measures

Eleven

Early detection of population declines: High power of genetic monitoring using effective population size estimators

Tiago Antao, Andrés Pérez-Figueroa, Gordon Luikart

Abstract

Early detection of population declines is essential to prevent extinctions and to ensure sustainable harvest. We evaluated the performance of two N_e estimators to detect population declines: the two-sample temporal method and a one-sample method based on linkage disequilibrium (LD). We used simulated data representing a wide range of population sizes, sample sizes, and number of loci. Both methods usually detect a population decline only one generation after it occurs if N_e drops to less than approximately 100, and 40 microsatellite loci and 50 individuals are sampled. However, the LD method often outperformed the temporal method by allowing earlier detection of less severe population declines (N_e approximately 200). Power for early detection increased more rapidly with the number of individuals sampled than with the number of loci genotyped, primarily for the LD method. The number of samples available is therefore an important criterion when choosing between the LD and temporal methods. We provide guidelines regarding design of studies targeted at monitoring for population declines. We also report that 40 SNP (single nucleotide polymorphism) markers give slightly lower precision than 10 microsatellite markers. Our results suggest that conservation management and monitoring strategies can reliably use genetic based methods for early detection of population declines.

11.1 Introduction

Managers of threatened populations face the challenge of early and reliable detection of population declines. Maintenance of large populations and associated genetic variation

is important not only to avoid population extinction but also because loss of genetic variation affects the adaptation capability of a population. Timely detection of populations that have suffered a decline will allow for a broader and more efficient range of management actions (e.g., monitoring, transplanting, habitat restoration, disease control, etc.) which will reduce extinction risks.

Genetic methods can be used to estimate effective population size (N_e) and monitor for population declines (Leberg, 2005). N_e is widely regarded as one of the most important parameters in both evolutionary biology (Charlesworth, 2009) and conservation biology (Nunney and Elam, 1994; Frankham, 2005).

The most widely used genetic method for short-term (contemporary) N_e estimation (Krimbas and Tsakas, 1971; Nei and Tajima, 1981; Pollak, 1983) is based on obtaining two samples displaced over time (generations) and estimating the temporal variance in allele frequencies (F) between them. Luikart et al. (1999) demonstrated that the temporal method was far more powerful than tests for loss of alleles or heterozygosity for detecting population declines. However little is known about the relative power of other N_e estimators for early detection of declines. Single sample methods based on linkage disequilibrium (LD), have been proposed (Hill, 1981; Waples, 2006) and have been compared to the temporal method for equilibrium (i.e., stable population size) scenarios (Waples and Do, 2010). Methods to estimate long-term effective size (Schug et al., 1997) are by definition not generally applicable to the problem of detecting a recent sudden change in effective size.

Here we evaluate and compare the power, precision and bias of both methods used to estimate N_e for early detection of population declines. We use simulated datasets from population declines with a wide range of bottleneck intensity, sample size and number of loci. We simulate both highly polymorphic loci (microsatellites) and biallelic loci (single nucleotide polymorphisms, SNPs). We also study, to a smaller extent, a more recent temporal method based on likelihood (Wang, 2001; Wang and Whitlock, 2003).

We address important questions posed by conservation biologists such as, “To establish a monitoring program, how many individuals and loci are needed to detect a decline to a certain N_e ?”, “How many SNPs are required to achieve sensitivity equal to microsatellites to estimate N_e and detect declines?”, “How many generations after a population decline will a signal be detectable?”, “What is the probability of failing to detect a decline (Type II error)?”.

11.2 Methods

We conducted simulations using the forward-time, individual based simulator SimuPOP (Peng and Kimmel, 2005). The default scenario was based on a constant size population of $N=600$ run until mean heterozygosity reached approximately 0.8 (10 generations) split into a number n of subpopulations ($n = 1, 2, 3, 6, 12$) without any migration. This in

practice simulates a bottleneck (with the exception of $n=1$). The average sex ratio was 1 with random mating. This approximates $N_c = N_e$. Each scenario was replicated 1000 times. For convenience, the census size before the bottleneck will be called N_1 and after will be labelled N_2 . Unless otherwise stated, when referring to equilibrium scenarios, we are mainly concerned with a population of constant size (e.g. $N_1 = N_2$ above).

The genome simulated includes, 100 neutral, independent microsatellites initialized with a Dirichelet distribution (10 alleles exhibiting a mean of 8 at the generation before the bottleneck) and no mutation.

We also compared and evaluated both methods according to:

1. Sensitivity to mutation rate. We used the K-allele model (Crow and Kimura, 1970) with 10 alleles and a relatively high mutation rate of 10^{-3} typical of some microsatellites (Ellegren, 2004).
2. Usage of SNPs. We conducted simulations using genomes with 100 physically unlinked SNPs initiated from a uniform distribution.
3. Sensitiveness to initial population size. We used different initial population sizes (2400, 1200, 600, 300) all bottlenecking to an N_2 of 50.
4. Benefits of using additional loci versus additional samples. While for equilibrium scenarios adding more loci is roughly equal to adding an equal proportion of individuals sampled (Waples and Do, 2010; Waples, 1989), we investigated if this symmetry holds under a population decline. We constructed a scenario with $N_1=300$ and $N_2=50$ and used different sampling strategies: 50 loci with 10 individuals and 10 loci with 50 individuals.

The simulation application saves for analysis all individuals in the generation exactly before the bottleneck along with 1, 2, 3, 4, 5, 10 and 20 generations afterwards. Each replicate is then sampled to study the effect of the sample size of individuals and loci. For N_e estimation we only study a single sub-population after each bottleneck to assure independence of all estimated values among replicates. We use for the number of loci 10, 20, 40 (and 100 for SNPs) and for the number of individuals 25 and 50. For each simulation replicate the following statistics are computed under different sampling conditions using Genepop (Rousset, 2008) through Biopython (Cock et al., 2009): F_{ST} (Weir and Cockerham, 1984), expected heterozygosity, and allelic richness.

To study the LD method each simulation replicate was analysed with the LDNe application (Waples and Do, 2008) which implements the bias correction (Waples and Gaggiotti, 2006) to the original LD method (Hill, 1981). Point estimates and 95% confidence intervals (parametric) are stored using only alleles with a frequency of 2% or more which is reported to provide an acceptable balance between precision and bias (Waples and Do, 2010) for the sample strategies tested.

For the temporal method we implemented the N_e estimator from Waples (1989) based on Nei and Tajima (1981):

$$\hat{N}_e = \frac{t}{2 \left[\hat{F}_k - \frac{1}{2S_0} - \frac{1}{2S_t} \right]} \quad (11.1)$$

Where t is the time between generations, S_0 is the sample size at the reference, pre-bottleneck point and S_t at the post-bottleneck generation being considered. The F_k estimator was implemented for each locus (l) as (Krimbas and Tsakas, 1971; Pollak, 1983):

$$\hat{F}^l = \frac{2}{K-1} \sum_{i=1}^K \frac{(f_{ri} - f_{ti})^2}{f_{ri} + f_{ti}} \quad (11.2)$$

Where K is the number of alleles at the current loci, f_{ri} is the frequency of allele i at the reference time and f_{ti} is the frequency of allele i at the current time. The generation before the bottleneck is used as the reference point to which all the other post bottleneck samples are compared. The F_k value used in the N_e estimator will be the weighted arithmetic mean of all locus F_k estimators, being the weight the number of alleles.

Confidence intervals on \hat{F} , which can be used to calculate the CI of \hat{N}_e , were computed as follows (Waples, 1989; Sokal and Rohlf, 1995; Luikart et al., 1999):

$$\alpha(1 - \alpha)CI \text{ for } \hat{F}_k^l = \left[\frac{n' \hat{F}_k^l}{\chi_{\alpha/2[n']}^2}, \frac{n' \hat{F}_k^l}{1 - \chi_{1-\alpha/2[n']}^2} \right] \quad (11.3)$$

Where n' is the number of independent alleles given by:

$$n' = \sum_{i=1}^l (K_i - 1) \quad (11.4)$$

Where K_i is the number of alleles of locus K .

We also studied a more recent version of a temporal based method, MLNE (Wang, 2001; Wang and Whitlock, 2003) which is based on likelihood estimation of effective population size. The number of cases studied was limited to only two bottleneck scenarios as the computational cost makes an exhaustive evaluation expensive.

The Coefficient of Variation (CV) is commonly used as a measure of precision and it is useful to compare results with theoretical expectations as these expectations hold for equilibrium. The CV for \hat{N}_e based on LD is (Hill, 1981; Waples and Do, 2010):

$$CV_{LD}(\hat{N}_e) \approx \sqrt{\frac{2}{n''}} \left[1 + \frac{3N_e}{S} \right] \quad (11.5)$$

Where n'' is:

$$n'' = \sum_{i=1}^{L-1} \sum_{j=i+1}^L (K_i - 1)(K_j - 1) \quad (11.6)$$

The CV provides a theoretical insight on other potential sources of lack of precision of the estimator: number of alleles and sample size are also expected to influence the precision of the estimator and most previous simulation studies of equilibrium report behaviours in line with theory. It is therefore important to investigate if qualitative and quantitative results hold for bottleneck cases.

The CV of the temporal estimator was presented in Pollak (1983):

$$CV_T(\hat{N}_e) \approx \sqrt{\frac{2}{n'}} \left[1 + \frac{2N_e}{tS} \right] \quad (11.7)$$

Where t is the time number of generations between samples and S is the sample size. The temporal based estimator has another expected source of imprecision: the temporal distance between samples.

We evaluated performance of both methods from three different perspectives:

1. *Detection* of a decline from the pre-bottleneck effective population size, e.g., to detect if the N_e (point estimate) is below $0.8 \cdot N_1$. This is similar to bottleneck tests (e.g. Cornuet and Luikart (1996)), as we are not concerned with the ability to approximate N_2 , only to detect if the population size decreased. The value chosen is arbitrary, but close to, and a function of N_1 .
2. *Approximation* of an effective population size that has declined closer to N_2 than to N_1 . Here we try to understand if, adding to the previous ability to detect a decline, an estimator (point estimate) can approach the new effective size. For instance if there is a bottleneck of $N_1=600$ to $N_2=50$, we want to study the ability of estimators' point estimate to be below 75, which is 50% above N_2 . This quantifies the ability to detect a change in N_e , but will not distinguish between an unbiased estimate of N_2 and downward bias one.
3. *Estimation* of N_2 with low bias and high precision and reliable confidence intervals. Most studies of equilibrium scenarios (stable population size) are of bias and precision and thus most comparable with this third perspective (e.g. England et al. (2006); Wang and Whitlock (2003)).

The three perspectives above are presented as they might be useful in different situations: A practical research question might need only to detect that a population is declining (detection perspective) or it might require that a certain conservation threshold (e.g., $N_e < 100$) has been passed (approximation perspective) or, still, a precise and

unbiased estimation of population size (estimation perspective). The first two perspectives are not applicable in equilibrium settings, but provide insights needed for practical conservation applications.

Methods for detection of population decline are reliable if, when there is no decline, the method does not erroneously suggest one (Type I error). This effect is especially important with N_e estimators as their variance is known to increase with increasing real N_e . As such we also assess how often each estimator detects a decline when there is none (false positive rate).

When characterizing the distribution of \hat{N}_e across simulations, we use mainly box plots. Box plots show the median, 25th and 75th percentiles, the lowest datum within 1.5 of the lower quartile and the highest datum within 1.5 of the quartile range. Other measures like e.g. mean squared error, can be calculated from the supplementary material.

We supply, as supplementary material, the distribution of N_e estimates (point, upper and lower CI) according to the boundaries specified in the perspectives above (i.e., the percentage of estimations which fall above N_1 , $0.8N_1$, $1.5N_2$, $0.5N_2$ or below $0.5N_2$ for all scenarios studied for the first five generations following the population decline. We also supply a set of standard population genetics statistics for (F_{ST} , expected heterozygosity and allelic richness) starting from the generation before the bottleneck up to 50 generations after. This material can be loaded in standard spreadsheet software for further analysis. Furthermore we also include an extensive number of charts covering all statistical estimators for all scenarios studied. Supplementary material is made available on <http://popgen.eu/ms/ne> or on the supplemental CD.

11.3 Results

With a fixed initial effective population size (N_1) of 600 and a population decline to an N_2 of 50, we could detect a reduction of \hat{N}_e from the original N_1 (detection perspective) after only one generation in 80% or more cases for each method when sampling just 25 individuals and 20 microsatellite loci. For an N_2 of 100 the temporal method detected the decline only after a few generations or by using more samples or loci, while the LD based method still immediately detects a decline with just 25 individuals and loci. If N_2 only drops to 200, the LD method will have still have power above 80% with 20 loci and 50 individuals at the first generation after the decline. Generally, the ability to detect a decline decreases for higher N_2 for both estimators as expected from the CV (Waples and Do, 2010) of both estimators.

Both methods were able to approximate N_2 (i.e. compute an estimation below $1.5N_2$) at generation 2 with a severe bottleneck of $N_2 = 50$ if 50 individuals were sampled. However the temporal method never had power above 80% for less severe

bottlenecks ($N_2 = 100$) in the first two generations. The power to detect an $N_e < 1.5N_2$ (approximation perspective) is presented in Figure 11.1.

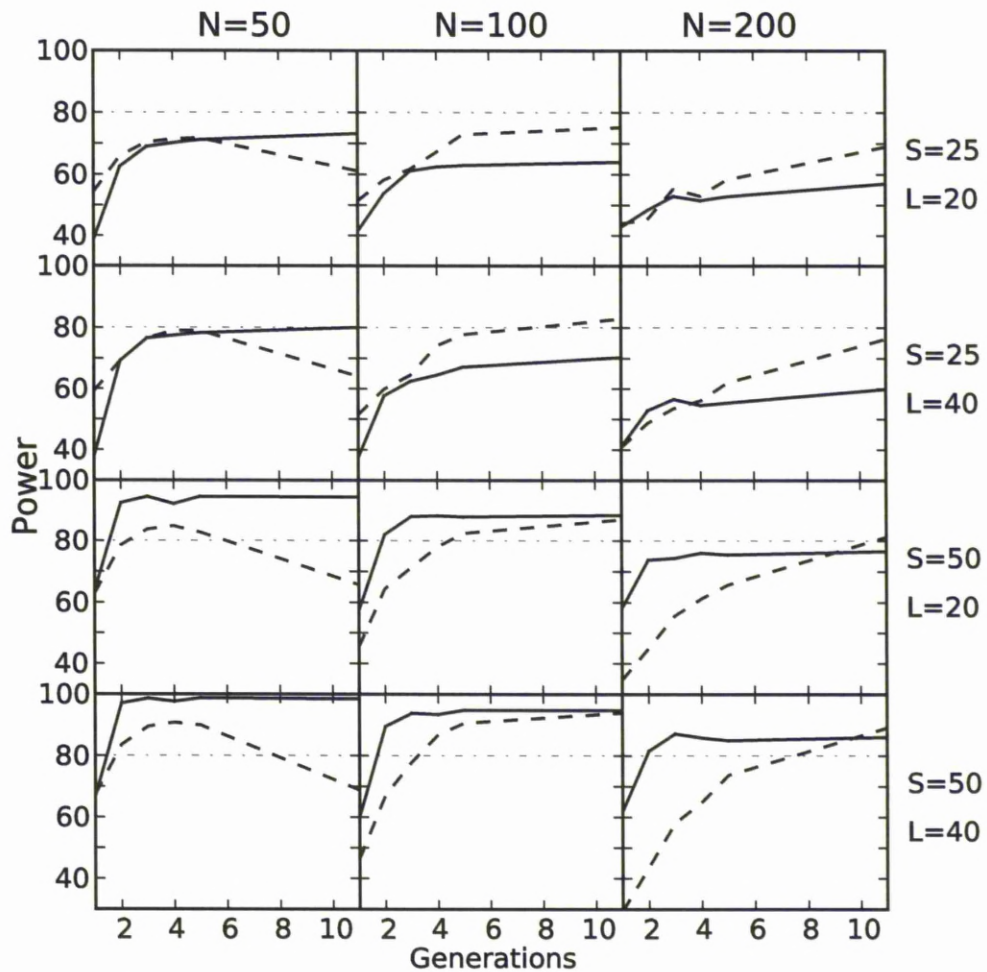


Figure 11.1: Power to detect that the effective population size (point estimate) is below 150% N_2 . The LD based method is shown as a solid line and the temporal method with a dashed line. The horizontal dotted line is the 80% power threshold. Each column comprises a different N_2 (50, 100, 200). The first row depicts 25 individuals and 20 loci; the second row 25 individuals and 40 loci; the third 50 individuals and 20 loci; the fourth 50 individuals and 40 loci.

As theoretically expected, power for early detection of a decline increases if more individuals are used. However, the following deviations from expectations (Waples, 1989; Waples and Do, 2010) are observed and further investigated in the Discussion section:

1. For the temporal method and for an N_2 of 200, power decreased slightly with more samples.

2. Increasing the number of individuals sampled is more beneficial for both methods than increasing the number of loci. This effect is more noticeable with the LD method.

For the estimation perspective (i.e. low bias and small confidence intervals; see Methods), our bias and precision analysis showed that the temporal method has lower precision and, with larger N_2 , higher bias upwards than the LD method. With a very low number of individuals, the LD method is biased upwards (consistent with England et al. (2010)) and less precise than the temporal method in line with the effect presented above (Figure 11.2).

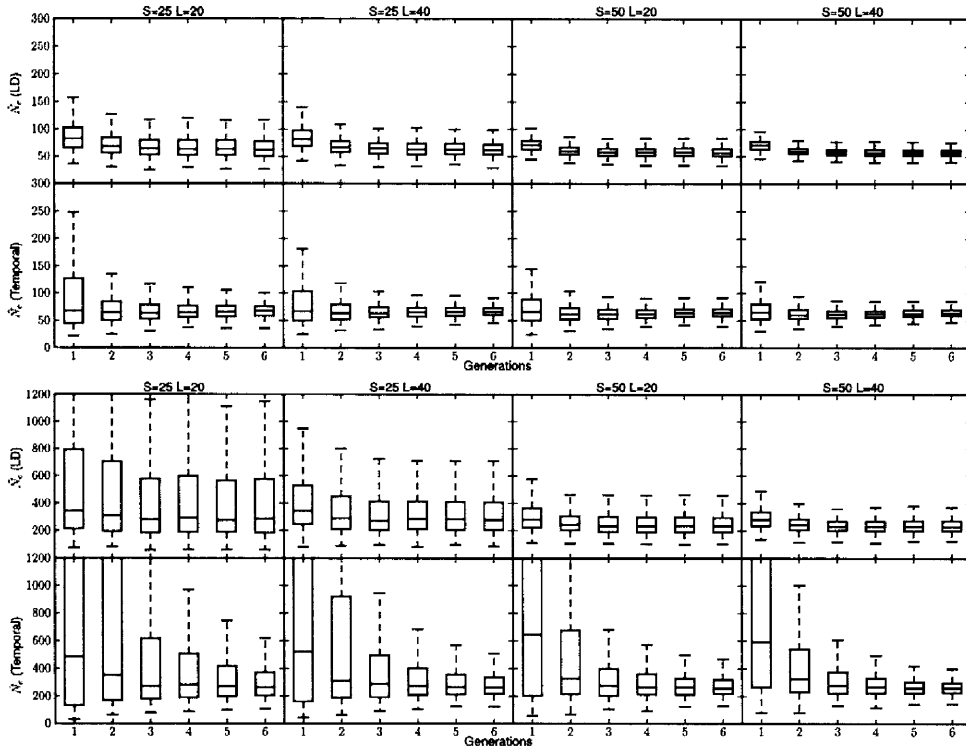


Figure 11.2: Boxplot charts of both the LD and temporal point estimates up to six generations after a bottleneck with $N_1=600$. The top chart reports a $N_2=50$, and the bottom chart a $N_2=200$. Different sampling strategies are shown on each column. On each chart, the top row depicts the LD method while the bottom row is the temporal method.

MLNE did not perform better than the original moments-based temporal method. We used MLNE with two bottleneck scenarios (N_2 of 50 and 200) and a sampling strategy using only two time points and MLNE never provided a reliable estimation even for large sample of 50 individuals and 40 loci. MLNE results were only usable with

3 samples in time but estimates were generally above N_2 in concordance with Wang (2001) which also reports over-estimation of N_e in non-equilibrium scenarios (further details and an estimation perspective with MLNE are presented in the supplementary material).

In order to understand the relative benefit of increasing the number of loci versus increasing the sample size, we simulated bottlenecks with an $N_1 = 300$ and a $N_2 = 50$ using two radically different sampling strategies: One maximizing the number of individuals (i.e., using a sample size equal to N_2) but using only 10 loci and another using 50 loci but only 10 individuals. The scenario with 5 times more individuals than loci gave higher precision in both methods. This effect was more pronounced with the LD method as both bias and precision are affected during all the initial five generations (Figure 11.3). The temporal method is mainly affected in precision, and only in the two initial generations for the scenario studied.

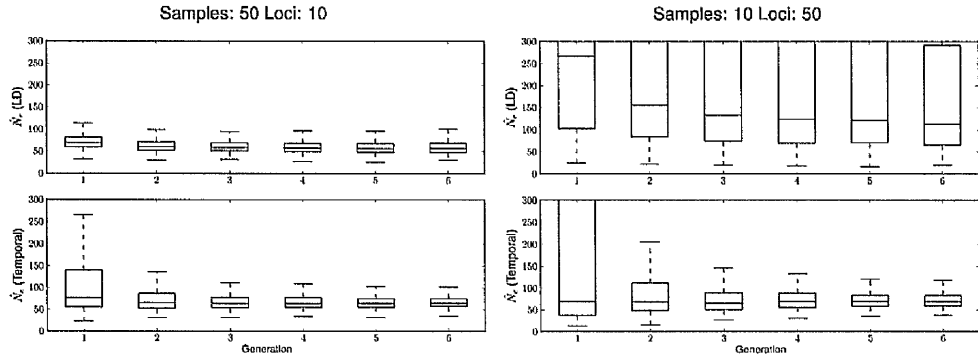


Figure 11.3: Boxplot of the \hat{N}_e during the first five generations of a bottleneck from $N_1=300$ to $N_2=50$. The left column depicts a sample size of 50 and 10 loci and the right column 10 individuals and 50 loci. Top row is the LD method and bottom row, the temporal method.

We also studied the behaviour of confidence intervals for both estimators. The upper confidence interval of the temporal method is often far higher than the initial population size during the initial bottleneck generations in most scenarios. This effect rarely occurs with the LD method: only on the very first generation and for high values of N_2 (Figure 11.4).

The usefulness of any estimator to detect a decline can be jeopardized by false positives, i.e., detection of a reduction in N_e when none occurred. We assessed the false positive rate for both estimators, i.e. with a true N_e of 600 (Figure 11.5). The LD-based method lower quartile of estimates was always above 400, whereas the lower quartile of point estimates for the temporal method approaches 200 when the sample size is only 25 individuals. For a sample size of 50 the LD method point estimates were normally above

500 whereas the temporal method point estimates were occasionally only approximately 100 even though the true N_e was 600.

We also studied how the pre-bottleneck size (N_1) affects the behaviour of the estimators. We simulated bottlenecks with different initial population sizes (initial $N_1=1200$, 600, 400, 300, 200 and final $N_2=50$, supplementary data). The LD method was little influenced by initial size, but the temporal method accuracy and precision decreased as N_1 decreased. This effect was mostly visible on the first generation after the decline, and disappears shortly after. This means that, adding to Type I errors which make methods less reliable to high N_1 , the temporal method also has precision problems with a lower N_1 .

We also quantified precision for bi-allelic markers (i.e., SNPs). Using 10 and 40 microsatellites and 40 and 100 SNPs a comparison among the distributions of the point estimates reveals results consistent with theoretical expectations (Figure 11.6). As an example, 10 microsatellite loci gave slightly higher precision than 40 SNPs: As the median allelic richness for the microsatellite scenarios after the bottleneck is 6 (supplementary material) the number of degrees of freedom (i.e. approximately the number of independent alleles) of the 40 SNPs scenario is smaller (20) than the 10 microsatellite scenario (50). The bias with SNPs is slightly lower probably because rare allele effects occurred less with bi-allelic markers we simulated. Type I errors also behave as expected, which for the sampling strategies shown and with equilibrium scenarios, gives not enough precision to differentiate between a Type I error and a real decline, again making Type I errors a fundamental consideration.

We also quantified the influence of mutation rate on the ability to estimate N_e . The number of new mutations is negligible in small populations over 1-10 generations even with high mutation rates. As an example, for an N_e of 100 and a relatively high mutation rate of 0.001 the expected number of new mutations per generation per locus would be 0.2 ($2 * N_e \mu$). Simulation results show negligible effect (supplementary material).

11.4 Discussion

Our results show that early detection and reliable size-estimation of population declines is increasingly possible using genetic monitoring and estimators of effective population size. Early detection is important as it allows for rapid management actions to avoid irreversible loss of genetic variation and increased risk of extinction due to genetic and demographic factors. Reliable estimation of N_e and the change in N_e is crucial in conservation biology but also in studies of evolution and ecology, for example to quantify bottleneck size associated with founder events or colonization of new environments.

LD method

The one-sample LD method generally outperformed the two-sample temporal method by allowing earlier detection of less severe population declines ($N_2 > 100$) when using sample sizes of loci and individuals typical of studies today. Nonetheless, if the number of individuals sampled is low (≤ 25), the temporal method might be a better option, especially if multiple generations pass between temporal samples. Both methods were able to approximate the N_e of a bottlenecked population fairly quickly especially for N_e below 200, in most cases in less than three generations after the decline event. The generation number after the bottleneck might alter the relative performance of the estimators in a qualitatively meaningful way (e.g. bias is in an opposite direction for each estimator immediately after the bottleneck versus several generations after). Here we are concerned with early detection, thus we note that some conclusions here might not hold if the generation gap is above 5-10 generations, which we did not study.

Temporal method

Experimental design (e.g. for a monitoring program) is more complex in the temporal method. Having two samples that are close temporally can yield relatively low precision (Wang and Whitlock, 2003) while having two samples that are separated by many generations, biases the estimate up-ward (Richards and Leberg, 1996; Luikart et al., 1999). This effect is easier to control in equilibrium scenarios as the underlying assumption of equilibrium would allow for some calibration of the distance between samples. But in non-equilibrium scenarios the uncertainty of a possible decline event between samples makes calibration less obvious.

For the temporal method and large N_2 , power declines as more individuals are sampled. This counter intuitive result can be attributed to two simultaneous causes: (i) Sampling more individuals raises the probability of increasing the number of rare alleles detected. Rare alleles are known to bias upward the temporal method (Turner et al., 2001), while the LD-based method includes an explicit correction (Waples and Gaggiotti, 2006). (ii) On the other hand, a smaller decline (higher N_2) will purge rare alleles more slowly. The precision of the temporal method is increasing with the number individuals sampled, but it is increasing towards an upward bias result, whereas the power definition used (relevant for the detection of a population decline) is concerned with detecting a value lower than a certain threshold. A correction to the temporal method (Jorde and Ryman, 2007) to deal with upward bias does exist, but it is known to have a larger standard deviation than the original method (which is already large for the initial generations after the decline).

The MLNE method did not provide any improvement with only two sample time points. If more time points are available then MLNE might provide more reliable results, but in a context of early and timely estimation of a population decline this requirement

for extra data might lower the usefulness of the method. Further research in MLNE is impaired by its computational cost (a study of the MLNE cost is available in the supplementary material).

Confidence intervals for the LD method are generally much tighter than the temporal method even after the first 3-5 generations. While the interpretation of confidence intervals for both methods is not always straightforward (an exhaustive discussion can be seen in (Waples and Do, 2010)) and its relevance is open to discussion, it is clear that there is a qualitative difference between estimators for early detection: The upper confidence interval for the temporal method often includes very high values, this is mainly caused by the known behaviour of the estimator to have poor precision for samples with only a few generations between samples.

Equations 11.1 and 11.2 show that the reference (i.e. before decline) and current (i.e. after decline) time are commutative in the temporal method. Our results show that the temporal method, when using a sample from before the bottleneck and another for after, tends towards the lowest value. This fortunate effect is fundamental in order to use the temporal approach to detect a decline. If the method did not approximate the lowest value then the first reference sample that could be used would be one immediately after the bottleneck, therefore delaying any estimation of a decline.

Effect of pre-bottleneck size

The temporal method is also sensitive to the pre-bottleneck size for the estimation of N_e after decline. The more similar the size of the population before (N_1) and after (N_2) the decline the worse the temporal estimator performs. This has implications for the feasibility of genetic monitoring studies based on the temporal method: On one hand, as the pre-bottleneck size increases, the Type I error also increases, on the other hand the closer N_1 is to N_e , the larger the Type II error (i.e. failure to detect a decline). Therefore, while the LD method is only sensitive to large pre-bottleneck sizes, the temporal method is also sensitive to the relationship between pre- and post- N_e . Experimental design (monitoring) with the temporal method could be more complex because the effect is more noticeable for relevant values of N_e and the small sample sizes common in conservation genetics scenarios. This effect tends to disappear soon after the bottleneck, so it will depend on the specific case to determine if very early detection is needed or not as that will have implications in the applicability of the temporal method.

Importance of number of samples

When trying to detect population declines, adding more individuals appears more beneficial than adding more loci, especially for the LD method. While previous studies (Waples and Do, 2010; Waples, 1989) have suggested that, for equilibrium scenarios, adding more loci is roughly interchangeable with adding more individuals, that is not

the case when precise early detection of population decline is needed. This effect is unfortunate given that the ability to genotype more markers is fast increasing while sampling many individuals can be difficult for populations of conservation concern. When determining the feasibility of genetic monitoring strategies, researchers should be especially careful in determining if sampling of enough individuals at any point in time is feasible. As the temporal method is often less prone to this effect – in fact it might not even be affected at all as empirical analysis suggest (Palstra and Ruzzante, 2008) – if the ability to sample many individuals is low then the temporal method might be a better option. Further research is needed to formally characterize both estimators after a bottleneck, especially trying to understand how past history and current state influence precision and bias and why the benefits of adding loci and samples are not similar to non-equilibrium scenarios.

Single Nucleotide Polymorphisms (SNPs)

As expected, SNPs provide less precision and accuracy (per locus) than microsatellites for estimation of N_e . Both methods depend on the number of independent alleles for a precise estimate in equilibrium populations, therefore the expectation is that using bi-allelic loci will provide lower precision compared to microsatellites; this also appears true for declining populations. Nonetheless, the bias with SNPs is slightly lower (probably because rare allele effects occur less with bi-allelic markers we simulated). Again, the rate of false positives becomes a fundamental consideration, impairing the ability to detect a decline (supplementary material). Further research is needed to quantify effects of different numbers of alleles (e.g., replacing microsatellites with a higher number of SNPs) in non-equilibrium scenarios, especially as we have demonstrated that, contrary to equilibrium scenarios, increasing the number of loci and sample size do not equally improve precision and accuracy.

Assumptions

Three assumptions of our work deserve mention and future research. First, future research is needed to quantify the effects of violating the assumption of no migration, and to develop methods to jointly estimate N_e and migration that is generalizable over a range of metapopulation models. Methods have been proposed (Vitalis and Couvet, 2001; Wang and Whitlock, 2003) to jointly estimate migration and N_e but have not been thoroughly evaluated or are not highly generalizable (e.g. beyond equilibrium populations or continent-island metapopulation systems). Another important assumption is non-overlapping generations. Most methods has not been extended to (or evaluated for) species with overlapping generations or age structure (but see Jorde and Ryman (1995) and Waples and Yokota (2007)). A third important assumption that requires thorough evaluation is mating system or behaviour which could bias the LD method, e.g. if a

dating system generates LD. Further research is needed to assess the importance of the issues presented above before applying the N_e estimators to scenarios mentioned above.

Type I error rate

False positives (i.e., Type 1 errors) are a concern when designing a study or monitoring program to detect a population decline, because false positives can lead to the waste of conservation resources on populations not actually declining. The temporal method is arguably more prone to false positive detection than the LD method. The need for more individuals or markers can sometimes be justified not by bias and precision in estimating post-bottleneck sizes but mostly by the need to avoid false positives, as N_e estimators are less precise with the larger, pre-decline, real N_e . This false positive effect might be less important in some conservation cases where the original population is known to be very low even pre-bottleneck. As in some conservation management cases the consequences of acting when there is no need is normally much smaller than the cost of not acting when there is a need (i.e., Type II error), e.g. to avoid extinction, a somewhat high probability of false positives might, in any case, be acceptable although this will vary from case to case.

Other methods

Several other methods to estimate N_e have been proposed (Tallmon et al., 2008; Nomura, 2008) and a comprehensive comparison of performance would be useful. Likelihood based methods (like MLNE) are expected to be computationally intensive making comprehensive studies difficult as the computational cost to conduct a large number of simulations and posterior evaluation could be prohibitive. This questions the practical applicability of computationally intensive methods as comprehensive evaluations of performance and reliability will require vast amount of computational resources. Thus evaluation of performance often will be, in practice, limited to a small number of scenarios. Approximate Bayesian and summary statistic methods including multiple summary statistics (e.g. both temporal F and LD) could greatly improve precision and accuracy of N_e estimators (Tallmon et al., 2008; Luikart et al., in press), especially as large population genomic data sets become common making likelihood-based methods even more computationally demanding to evaluate (Luikart et al., 2003).

Conclusion

Early detection of population declines is increasingly feasible with the use of genetic monitoring based on effective population size estimators. If the number of samples is sufficiently high, LD based method is arguably more powerful and better suited for monitoring to detect declines because it is less prone to Type I errors, has tighter

confidence intervals, and is more flexible with regards to designing different experimental design strategies. Nonetheless it is important to further research the behaviour of both estimators under an even broader set of realistic scenarios, e.g. with age structure or migration, and to understand if variations of the temporal method (Jorde and Ryman, 1996) or LDNe allow for earlier and more precise estimation of effective population size in decline populations. Both methods along with others (e.g. loss of alleles) should often be used when monitoring in order to gain a better understanding of the causes, consequences and severity of population declines (Luikart et al., 1999).

As the precision of both estimators requires the true effective population size to be relatively small, their use is currently limited to scenarios in conservation biology and perhaps studies of the ecology and evolution in small populations. For instance, they cannot be used to conduct reliable genetic monitoring when the effective size remains larger than approximately 500 to 1000 unless perhaps hundreds of loci and individuals are sampled and/or improved estimators are developed.

Simulation evaluations of new statistical methods and increasing numbers of DNA markers makes management and genetic monitoring increasingly useful for early detection of population declines, even with non-invasive sampling of elusive or secretive species. These results are encouraging and contribute to the excitement and promise of using genetics in conservation and management.

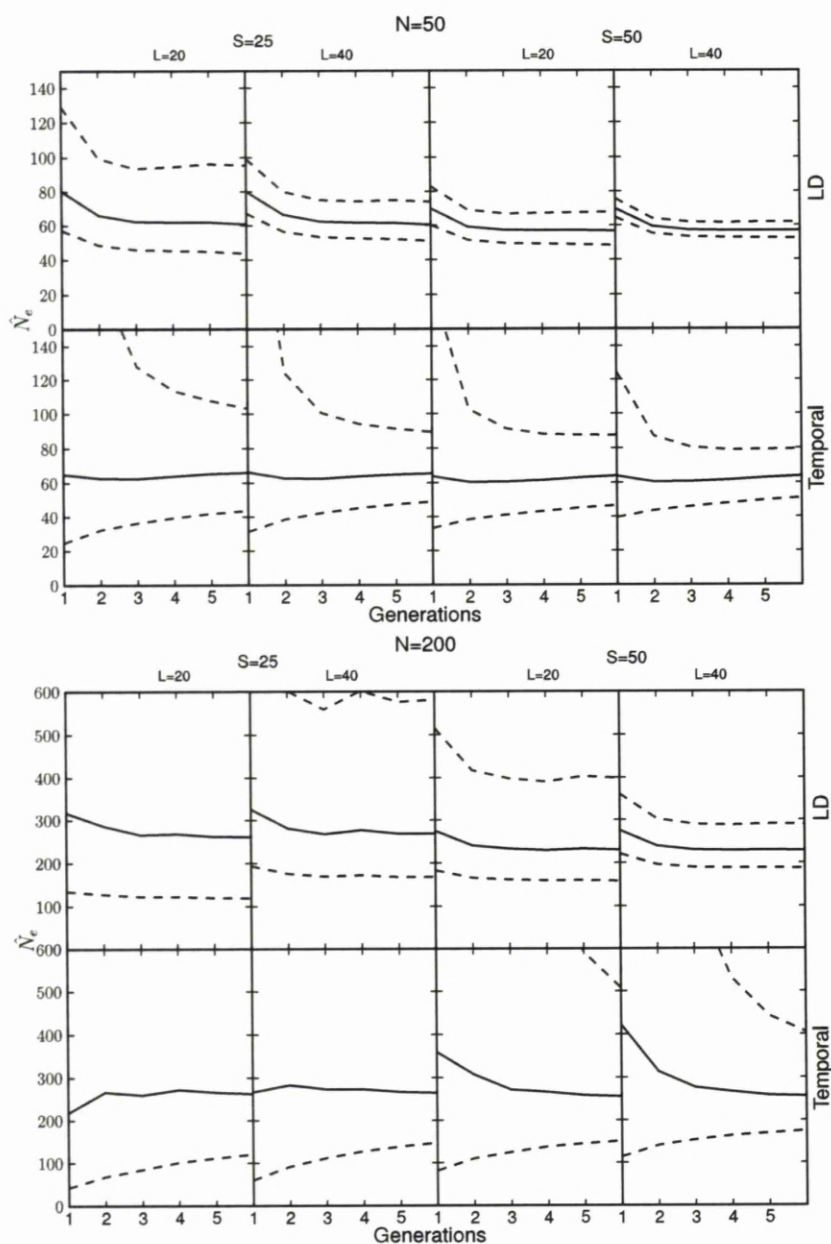


Figure 11.4: Harmonic mean of \hat{N}_e (solid line) and 95% confidence intervals of 1200 post-bottleneck replicates (dashed lines) for both methods for two bottleneck scenarios and four sampling strategies all with $N_1 = 600$. The first chart reports a $N_2 = 50$ and the second a $N_2 = 100$. Different sampling strategies are shown on each panel from left to right: 25 individuals and 20 loci on the first, increasing to 40 on the second; the third shows 50 individuals and 20 loci increasing to 40 loci on the far right.

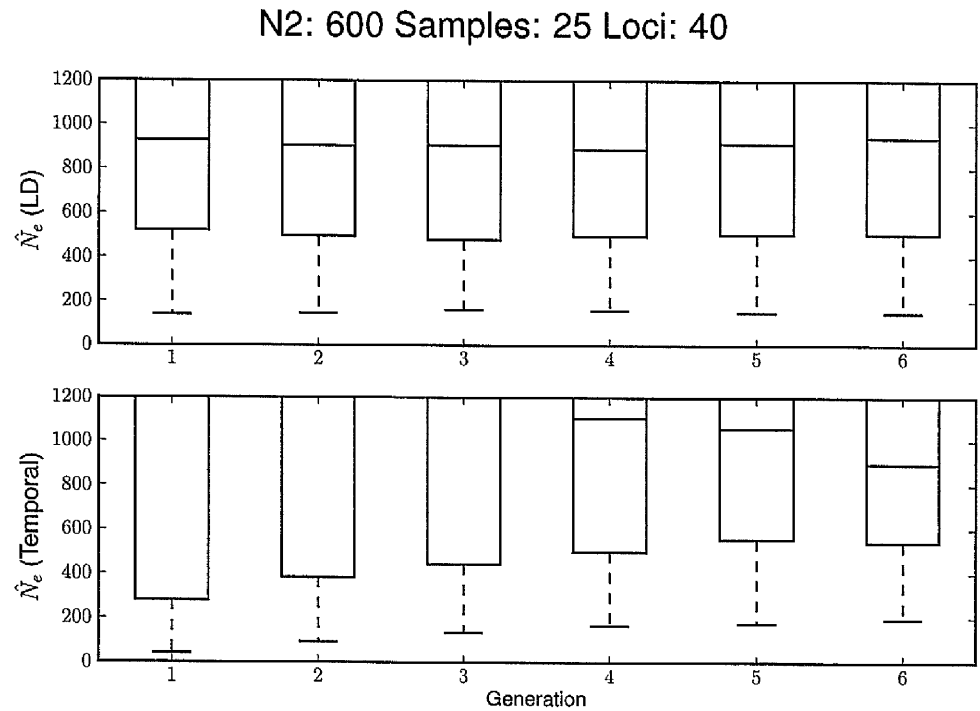


Figure 11.5: Boxplot of the distribution of point estimates for both estimators under equilibrium ($N=600$) with 25 samples and 40 loci zooming in the area relevant for type I error detection. Estimates are biased high because the noise from sampling is often greater than the signal from drift or the number of parents.

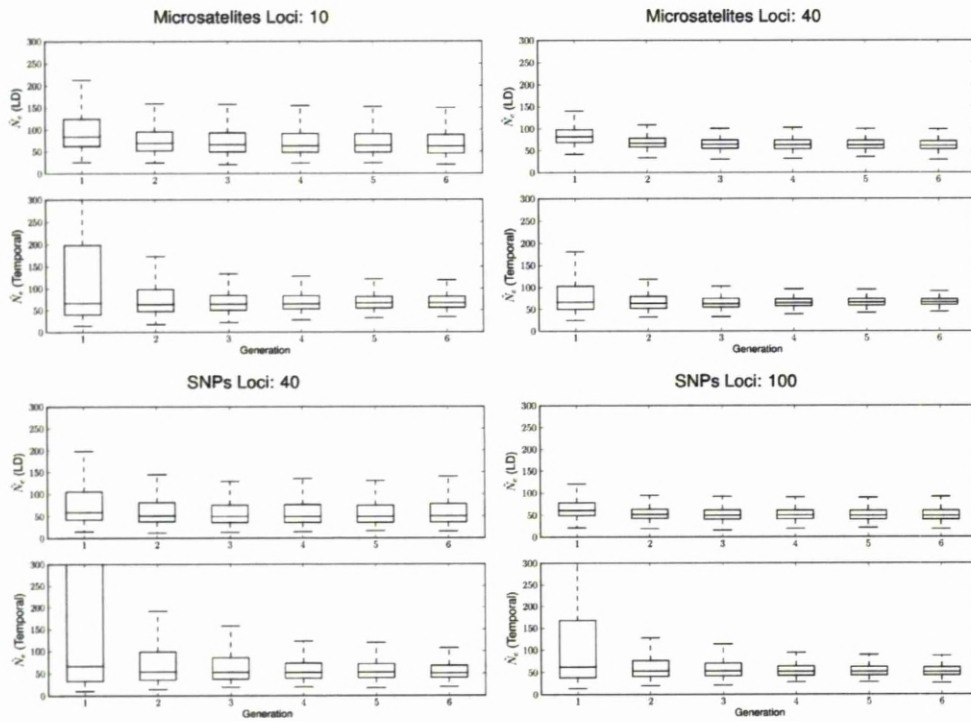


Figure 11.6: Boxplot of the \hat{N}_e of both estimators during the first five generations of a bottleneck from 600 to 50 with a fixed sample size of 25. The bottom row depicts SNPs, left 40 loci and right 100 loci. The top row depicts microsatellites, left 10 loci and right 40 loci.

Twelve

Estimating effective population size of disease vectors: a critical assessment of applications and performance

Tiago Antao, Ian M. Hastings,
Gordon Luikart and Martin J. Donnelly

Abstract

Estimation of the contemporary effective population size (N_e) is increasingly conducted for insect disease vectors. It is used to assess changes in genetic diversity due to drift and to evaluate the impact of control measures like insecticide-treated bed nets (ITNs) or indoor residual spraying (IRS). We evaluated the performance of the most commonly used N_e estimator based on F-statistics which uses two temporally spaced samples and compared it to two recent estimators based on likelihood (another 2-sample temporal method) and linkage disequilibrium (a 1-sample method). We simulated large N_e with three different demographies based on realistic parameters for common vector species: a constant population size model, bottlenecks to simulate effects of control measures, and a novel model with sinusoidal demography to simulate populations whose size fluctuates seasonally with dry and wet seasons. Results show that the sample sizes common in empirical studies (~ 60 individuals and ~ 10 microsatellite loci) are not sufficient to estimate N_e with precision. We suggest that an increase in temporal spacing between samples (e.g. > 12 generations) and an increase to ~ 50 loci provide sufficient precision. The likelihood method outperforms the F-statistics moments estimator in most cases. In fluctuating populations the temporal and LD methods provide qualitatively different estimates. The LD method is sensitive to immediate demographic changes whereas the temporal methods provide an average over several generations. Each class of methods may have different applications as the LD method can detect seasonal patterns and control interventions whereas the temporal estimators quantify trends over several generations.

12.1 Introduction

The effective population size (N_e) provides a measure of the rate of random genetic change in populations caused by genetic drift (Charlesworth, 2009), and the relative efficiency of natural selection in the face of drift. N_e is also a fundamental factor determining population viability as larger N_e entails greater population genetic variability which is paramount for species survival and adaptation. Estimation of N_e is increasingly conducted for insect disease vectors such as the *Anopheles* mosquito vectors of malaria and the fly *Glossina* vectors of African trypanosomiasis. Commonly the purpose of these studies is to determine whether insecticide-based control measures have successfully reduced the contemporary N_e (e.g. Wondji et al., 2005) or to investigate whether populations undergo seasonal contractions (e.g. Simard et al., 2000).

Indeed, one of the most widely used estimators of N_e across all taxa was developed to study the impact of insecticide resistance (Krimbas and Tsakas, 1971). This moments based method of contemporary N_e estimation was further developed (Nei and Tajima, 1981; Pollak, 1983; Waples, 1989) and is based on obtaining at least two samples displaced over time (generations) and estimating the temporal variance in allele frequencies between them.

Luikart et al. (1999) demonstrated that this temporal method was far more powerful for detecting population declines than tests based on loss of alleles or heterozygosity for detecting population declines. Effective population size should not be confused with census size (N_c), i.e. the total number of individuals in a population at any given time. N_c is normally larger than N_e for wildlife species with $\frac{N_e}{N_c}$ ratios estimated around 0.10 (Frankham, 1995; Kalinowski and Waples, 2002), although there is only one known study for this ratio in parasite vectors (Solano et al., 2009) where ratios calculated exhibit a large variance.

The second most widely used class of estimators of contemporary N_e is the single-sample estimator based the linkage (gametic) disequilibrium (LD) method (Hill, 1981). Waples and Do (2010) showed that LDNe can provide precise estimates of N_e in constant-sized populations with non-overlapping generations by using 10–20 microsatellite loci (5–10 alleles/locus) and samples of at least 25–50 individuals, if the effective population size is less than approximately 500.

While the temporal estimator has been studied both with constant sized populations and bottlenecks (see e.g. Luikart et al., 1999; England et al., 2010; Antao et al., 2011) for low N_e , there has never been a critical assessment of its usage alone or in comparison to single sample methods (LDNe) in the context of vector biology. This assessment is increasingly needed given the importance of vector-borne diseases, increasing risks of emerging disease following environmental changes and the need to evaluate control interventions to reduce vector population size.

Standard assumptions for insect vector studies

The assumptions commonly made for vector population studies are that individuals are sampled without replacement prior to reproduction (Dyer et al. (2009) and others) which is plan II sampling of Waples (1989). Following the work of Lehmann et al. (1998) most of authors have take a conservative estimate of 12 generations per year for the African malaria vector *Anopheles gambiae* which allows estimates to be compared between studies. Researchers on South East Asian anophelines have used a value of 10 generations per year (Walton et al., 2000). Consequently the reported values may underestimate the real N_e if there are more generations per year. Authors also generally assume that allele frequency change is solely attributable to genetic drift (Dyer et al., 2009), whereas selection or sampling bias due to substructure could also lead to allele frequency changes. The final assumption is of constant population size whereas many insect vectors exhibit extreme seasonal variation in census size with populations increasing following the onset of favourable conditions. For tropical mosquito species this is often the onset of rainy season (Charlwood et al., 1995; Taylor et al., 1993).

Studies of N_e also use different sampling strategies with varying number of individuals sampled, number of loci, and especially the temporal spacing between samples. It is also assumed that there is independent sampling of individuals and independent (unlinked) loci (Dyer et al., 2009). The former assumption may be difficult to defend giving the increasing evidence of stratification in vector populations (Weetman et al., 2010). A representative sample of N_e studies in vectors is shown on Table 12.1. We will concentrate on studies reporting high N_e (i.e. above 100), because the behaviour of estimates with low N_e (like the values reported in Solano et al. (2009)) has been widely studied (e.g Tallmon et al., 2010; Waples and Do, 2010; Berthier et al., 2002).

Alternative estimators of contemporary N_e have been proposed, either (i) sophisticated versions of the temporal method using Maximum Likelihood (ML), such as the ones implemented in the MLNE (Wang, 2001; Wang and Whitlock, 2003) or TM3 (Berthier et al., 2002) applications or (ii) completely different approaches based on a single sample and linkage disequilibrium (LD) (Hill, 1981; Waples, 2006) as implemented in LDNe (Waples and Do, 2008). MLNE and TM3 were shown to have better performance than the standard temporal method for constant sized populations (Wang, 2001; Berthier et al., 2002) but doubts were raised if that was the case with MLNE in bottleneck scenarios (Antao et al., 2011). The LD method has been compared to the temporal method for constant population size scenarios (Waples and Do, 2010) and bottleneck detection (England et al., 2010; Antao et al., 2011).

Here we present a computational study, using individual-based forward-time population genetic simulations, evaluating several contemporary N_e estimators using realistic demographies for insect vectors. The rationale for this study is two-fold:

Species	Sample loci and individuals	Number of alleles	Temporal spacing	\hat{N}_E	Motivation	Publication
<i>A. gambiae</i>	11/55	K between 45 and 64	85, 98	4,258 to 6,359		Lehmann et al. (1998)
<i>A. arabiensis</i>	9/50	7	4, 9, 40	229 to 1046	Seasonal	Simard et al. (2000)
<i>A. gambiae</i>	12/55	NA	12, 24	1049, inf, 1457	DDT impact	Pinto et al. (2003)
<i>A. arabiensis</i>	12/55	8	5, 12, 16, 33	135 to 649	ITN	Wondji et al. (2005)
<i>G. palpalis</i>	12/35	10	23	229 to 1046		Dyer et al. (2009)

Table 12.1: Representative sample of empirical studies of contemporary N_e . The sampling strategy includes approximate number of loci and individuals sampled. The number of alleles is an average approximation of the reported value (with the exception of Lehmann et al. (1998) where the total number of independent alleles is reported). The sample spacing is in months.

1. We critically appraise existing published studies of N_e in insect vectors and determine whether
 - The number of individuals and loci are sufficient to provide an unbiased estimate of N_e with a reasonably narrow confidence interval.
 - The temporal spacing between samples spans enough generations to provide a accurate estimation of N_e .
 - Whether the temporal method will allow us to estimate the impact of vector control or seasonality on vector population size.
2. We also provide guidelines for future studies of N_e by
 - Investigating if recent approaches to estimate N_e perform better than the original moments based temporal method.
 - Studying the impact of realistic vector demographies on the estimators.
 - Suggesting sampling strategies and which estimator is most capable of providing sufficient precision when studying the impact of control measures and seasonality.

12.2 Methods

We start by presenting an overview of the standard temporal method, followed by an introduction to maximum likelihood temporal estimator and the LD method. We then describe the simulations and sampling strategies.

Moment-based F_k temporal method

For the temporal method we implemented the N_e estimator from Waples (1989) based on Nei and Tajima (1981) and Krimbas and Tsakas (1971):

$$\hat{N}_e = \frac{t}{2 \left[\hat{F}_k - \frac{1}{2S_0} - \frac{1}{2S_t} \right]} \quad (12.1)$$

Where t is the time between generations, S_0 is the sample size (number of individuals) at the reference, pre-bottleneck point and S_t at the post-bottleneck generation being considered. This is the estimator for plan II of Waples (1989) (sampling destructively before reproduction). Though in our simulations we sample non-destructively (plan I), the difference between estimators with high N_e (like the values simulated here) is expected to be low (Waples, 1989). The plan II estimator has been extensively evaluated and its usage will allow for comparative analysis, furthermore it is the estimator commonly used for vectors.

The F_k estimator is implemented for each locus (l) as (Pollak, 1983):

$$\hat{F}_k^l = \frac{2}{K-1} \sum_{i=1}^K \frac{(f_{ri} - f_{ti})^2}{f_{ri} + f_{ti}} \quad (12.2)$$

Where K is the number of alleles at the current loci, f_{ri} is the frequency of allele i at the reference time and f_{ti} is the frequency of allele i at the current time. The F_k value used in the N_e estimator will be the weighted arithmetic mean of all locus F_k estimators (equation 12.2), the weight being the number of alleles.

Confidence Intervals (CI) on \hat{F} , which can be used to calculate the CI of \hat{N}_E , were computed as follows (Waples, 1989; Sokal and Rohlf, 1995; Luikart et al., 1999):

$$\alpha(1-\alpha)CI \text{ for } \hat{F}_k^l = \left[\frac{n\hat{F}_k^l}{\chi_{\alpha/2[n]}^2}, \frac{n\hat{F}_k^l}{1 - \chi_{1-\alpha/2[n]}^2} \right] \quad (12.3)$$

Where $1-\alpha$ is the proportion of CIs containing the real N_e and n is the number of independent alleles given by:

$$n = \sum_{i=1}^l (K_i - 1) \quad (12.4)$$

Where K_i is the number of alleles of locus i .

The coefficient of variation (CV) of the temporal estimator was presented in Pollak (1983):

$$CV(\hat{N}_e) \approx \sqrt{\frac{2}{n}} \left[1 + \frac{2N_e}{tS} \right] \quad (12.5)$$

Where S is the number of individuals sampled. The CV, a measure of dispersion, suggests that precision is increased if either the spacing between samples or the number of alleles increases. The estimator is also expected to lose precision with large real N_e , as the actual value is in the numerator of the CV.

This moment-based estimator is known to be biased upwards (Waples, 1989; Berthier et al., 2002) and rare alleles are largely responsible for the overestimation, so we also pooled (binned) all alleles with frequency below 2% into a single class. Results from binning were compared with the standard (without binning) estimator.

Likelihood-based temporal estimator

Several estimators have been proposed that use temporal sampling and maximum-likelihood (ML). For example, the ML method by Berthier et al. (2002) has been used in Dyer et al. (2009) to estimate N_e in a *Glossina palpalis palpalis* population in Equatorial Guinea. ML based estimators should provide better precision than moment-based estimators because they use more information from the data (Edwards, 1972) and the ML method used in Dyer et al. (2009) has been shown to perform better with very low N_e (i.e. 20) and when dealing with rare alleles (Berthier et al., 2002). This multiallelic method is based on coalescent simulation. Like most ML methods, extensive testing is computationally costly in terms of time. Strictly speaking this method is Bayesian as a maximum N_e prior has to be supplied. This estimator is implemented in the TM3 application. The parameters for each estimate were extracted from 20,000 coalescent simulations where the N_e estimator prior was capped at 15,000. The following summary statistics are computed: mode and the 0.025 and 0.975 quantiles (giving a 95% support interval). There is no CV for this estimator, but the same variables that influence the moments-based estimator (real N_e , time between samples, number of alleles and sample size – equation 12.5) are expected to impact this estimator in qualitatively similar ways.

Linkage disequilibrium estimator

Linkage disequilibrium can be used to estimate effective population size as its magnitude is a function of N_e and sample size. (Hill, 1981) noted that the variance of LD estimates among loci is a function of the effective population size and proposed an estimator based on LD. The original estimator has been shown to be downwardly biased if the sample size is smaller than the true N_e (England et al., 2006) and a bias correction has been proposed (Waples, 2006). The LD method has one main clear advantage over temporal

approaches: it requires only a single sample. The LD method implemented in LDNe (Waples and Do, 2008) has been compared to the moment based temporal method for equilibrium (i.e., constant population size) scenarios (Waples and Do, 2010) and bottlenecks (Antao et al., 2011). Evaluations of performance are also given in Tallmon et al. (2010) and England et al. (2010). The CV for this estimator is (Hill, 1981; Waples and Do, 2010):

$$CV_{LD}(\hat{N}_e) \approx \sqrt{\frac{2}{n'}} \left[1 + \frac{3N_e}{S} \right] \quad (12.6)$$

Where n' is:

$$n' = \sum_{i=1}^{L-1} \sum_{j=i+1}^L (K_i - 1)(K_j - 1) \quad (12.7)$$

Both the coefficient of variation and computational studies suggest that, like the temporal method, the LD estimator has increased absolute precision for low real N_e and larger sample sizes (n' increases with both number of loci and alleles per loci).

Point estimates and 95% confidence intervals (parametric) are computed using only alleles with a frequency of 2% or more in order to correct for upward bias. This correction is reported to provide an acceptable balance between precision and bias (Waples and Do, 2010) for the sample strategies tested (when $S > 25$).

Simulations and demographies

We conducted simulations using the forward-time, individual based simulator simuPOP (Peng and Kimmel, 2005). Simulations were performed using a Wright-Fisher model with separate sexes, random mating (average sex ratio of 1) and discrete, non-overlapping generations. This makes $\frac{N_c}{N_e} \approx 1$. Each demographic scenario was replicated 1,000 times. Simulations had a burn-in phase of at least 10 generations in order to approximate mean observed heterozygosity with realistic values (below 0.8). Longer burn-in periods were also tested, but results were qualitatively unchanged. The genome simulated included 50 neutral, independent microsatellite loci initialised with a Dirichlet distribution (10 initial alleles per locus exhibiting a mean of 8 after burn-in, approximating the conditions in Table 12.1) and a mutation rate of 10^{-4} using a stepwise mutation model (Lehmann et al., 1998). Simulations with a larger number of starting alleles (up to 20) were also conducted. All simulation data was saved in the Genepop (Rousset, 2008) format and automatically processed using Biopython (Cock et al., 2009).

Three different demographies were tested: i) a standard demography of constant size, ii) a bottleneck potentially imposed by a transmission control measure such as ITNs or IRS and iii) a novel model where the population size varies with a cosine function in order to model vector seasonality. The constant scenario was run with a N_c of 200, 400, 800, 1000 or 2,000. The bottleneck scenario started with a N_c of 5,000 or 2,000 which

were then reduced ten-fold. The seasonal scenario was based on $A\cos(\frac{2t\pi}{12}) + B$ where t is the generation, and A and B are parametrised according to the demography. The function above implicitly defines a period of 12 generations, based on 12 generations per year for *A. gambiae*. For A and B we used the parameters $A = 500, B = 700$, making the minimum N_e of 200 and the maximum of 1,200.

To perform the N_e estimation we sampled 60 individuals per generation and 10 microsatellite loci, in line with existing studies. We also studied the impact of doubling the sample size (120 individuals) and tested different numbers of loci sampled (20, 50 and 100). For both temporal methods we tested different temporal distances between sampling ranging from 4 generations to 100 (this does not apply to the LD method as it is based on a single sample). Due to the extreme computational cost of the ML method we only studied 100 replicates (instead of 1,000) per scenario.

12.3 Results

Moment based temporal method in constant populations

Bias

For a time span of 4 generations, the true N_e is always below the lower quartile (the value that defines the 25% of lowest point estimates) of the distribution of 1,000 independent simulation-based point estimates, and the upper quartile (the value that defines the 25% of highest point estimates) is always more than 3 times the real N_e value. Figure 12.1 shows the distribution of point estimates for spans of 4, 12 and 24 generations using different sampling strategies. For N_e of 1,000 and 2,000 the lower quartile occasionally lies below the real value, but only due to increased imprecision. For time spans of 12 and 24 generations the median point estimate (among 1000 replicates) is always above the real value (i.e. there is upward bias), but in most cases the upper quartile is below 3 times the true value. Binning, i.e. pooling all alleles with a frequency below 2% in a single class in order to reduce upward bias is useful for loci with many alleles but rarely reduces bias substantially (results not shown).

Precision and confidence intervals

For an N_e of 2,000, the confidence intervals of point estimates is large and the harmonic mean of the upper confidence limit is always above 6,000 ($3N_e$) assuming that the time span between sampled generations is below 25 (Figure 12.2). The CI for an N_e of 1,000, for a typical sampling strategy with 60 individuals and 10 loci will also be above 3 times N_e for all time spans. For this sampling strategy, a time span of one year with *A. gambiae* is not enough to have a upper confidence limit below 3 times the N_e values simulated (including a N_e of 500). The upper confidence limit will also never be below 3 times the N_e for all sampling strategies if the time span between generations is below 8 irrespective of N_e . As expected, sampling more individuals and/or more loci

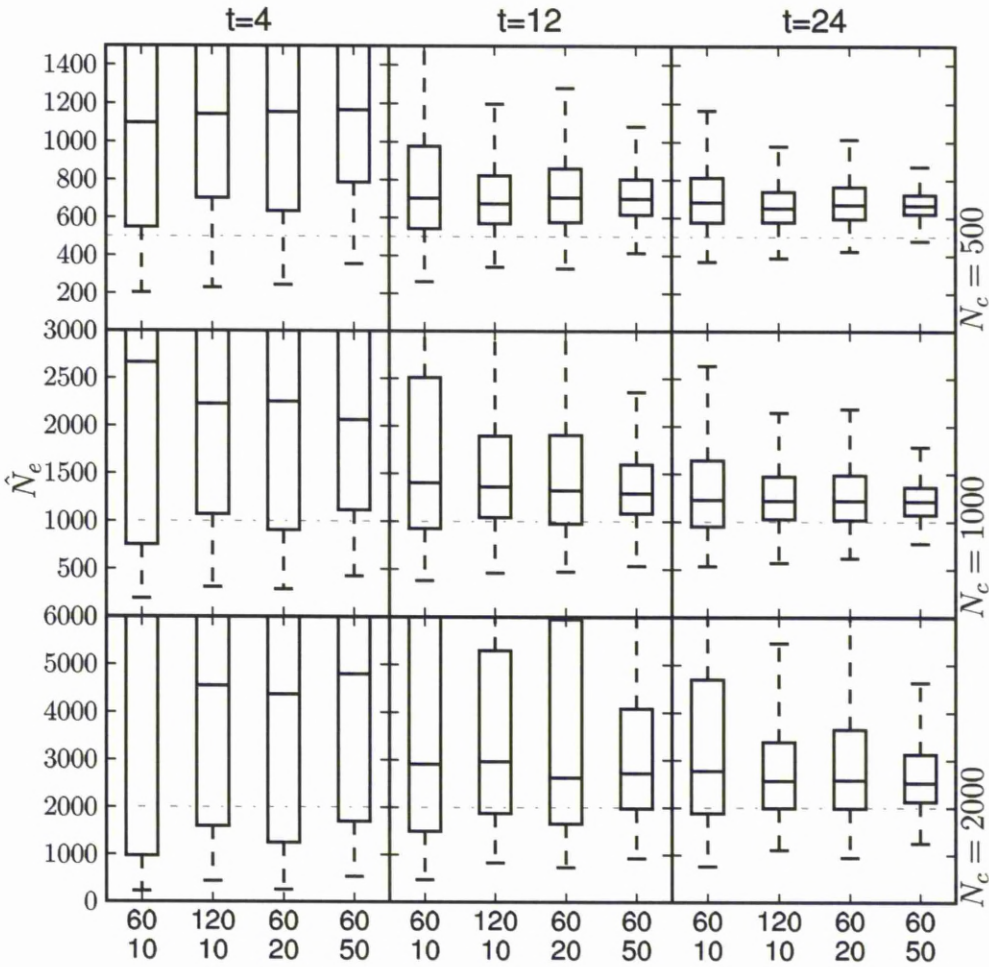


Figure 12.1: Boxplot charts of temporal point estimates obtained from moments-based F method for a time span between samples of 4, 12 and 24 generations. Four sampling strategies are considered: 60 individuals with 10, 20 and 50 loci and 120 individuals with 10 loci. The first row reports a constant $N_e=500$, the second line 1,000 and the third 2,000.

provides more precise estimates. Sampling more individuals appears to be slightly more informative than sampling more loci as the precision of sampling 60 individuals and 20 loci is slightly lower than the precision with 120 individuals and 10 loci. Figure 12.2 plots the harmonic mean for the point estimate and 95% confidence intervals with time spans between samplings up to 24 generations using different sampling strategies, estimations behind the horizontal line have more than 10% point estimates with infinite.

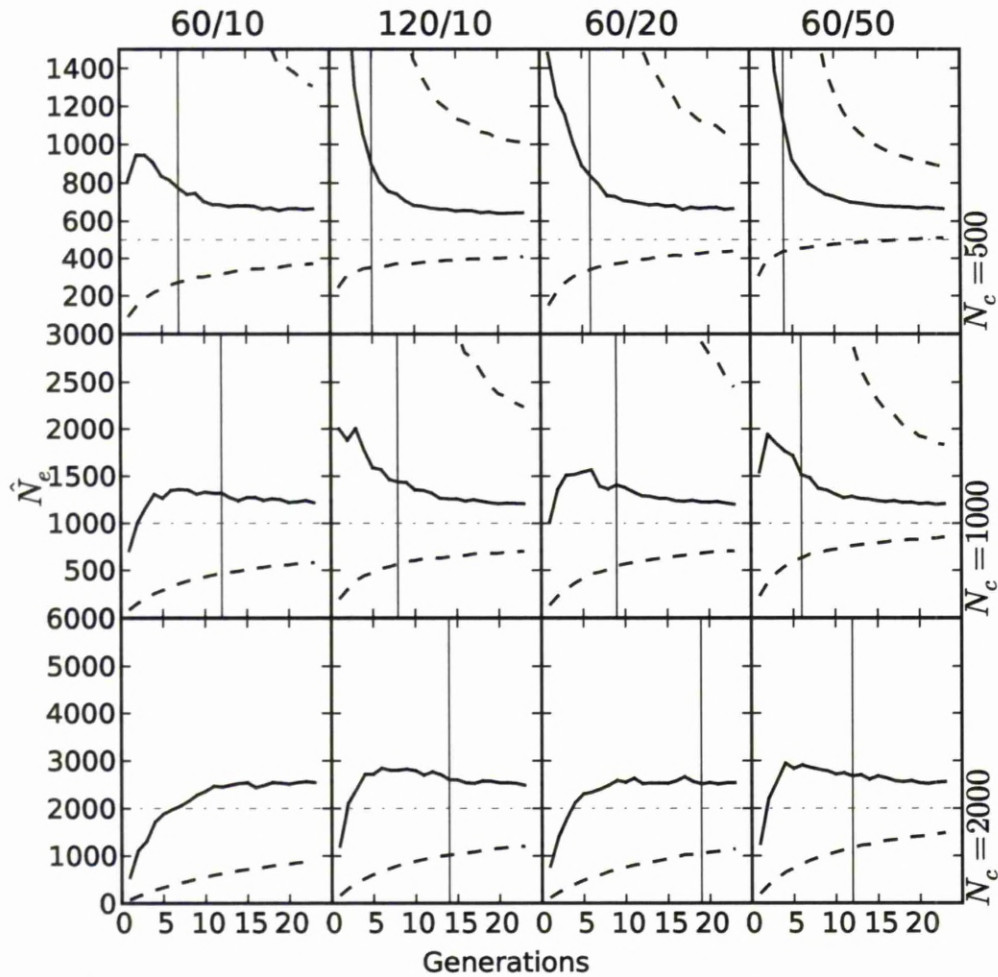


Figure 12.2: Harmonic mean of \hat{N}_E (solid line) and 95% confidence intervals (dashed lines) of 1,000 replicates for the moment-based temporal estimator for a time span between 1 and 24 generations. Four sampling strategies are considered: 60 individuals with 10, 20 and 50 loci and 120 individuals with 10 loci. The first row reports a constant $N_e=500$, the second row 1,000 and the third 2,000. Curves on the left of each vertical line have more than 10% of the 1000 point estimates equal to infinity.

Comparison of methods in constant-size populations

To compare the 3 methods used to estimate contemporary N_e , we show on figure 12.3 a box plot of the distribution of point estimates. For the temporal methods we include 2 time spans (4 and 24 generations). The ML method is more precise than the original moments based estimator assuming equal time spans with the exception of the typical sampling strategy and a time span of 4. A time span of 4 will always produce imprecise results unless the ML method is used with 60 individuals and 50 microsatellite loci.

While TM is more precise, it is also biased downwards with a real N_e of 1,000 as the upper quartile of point estimates is below 1,000.

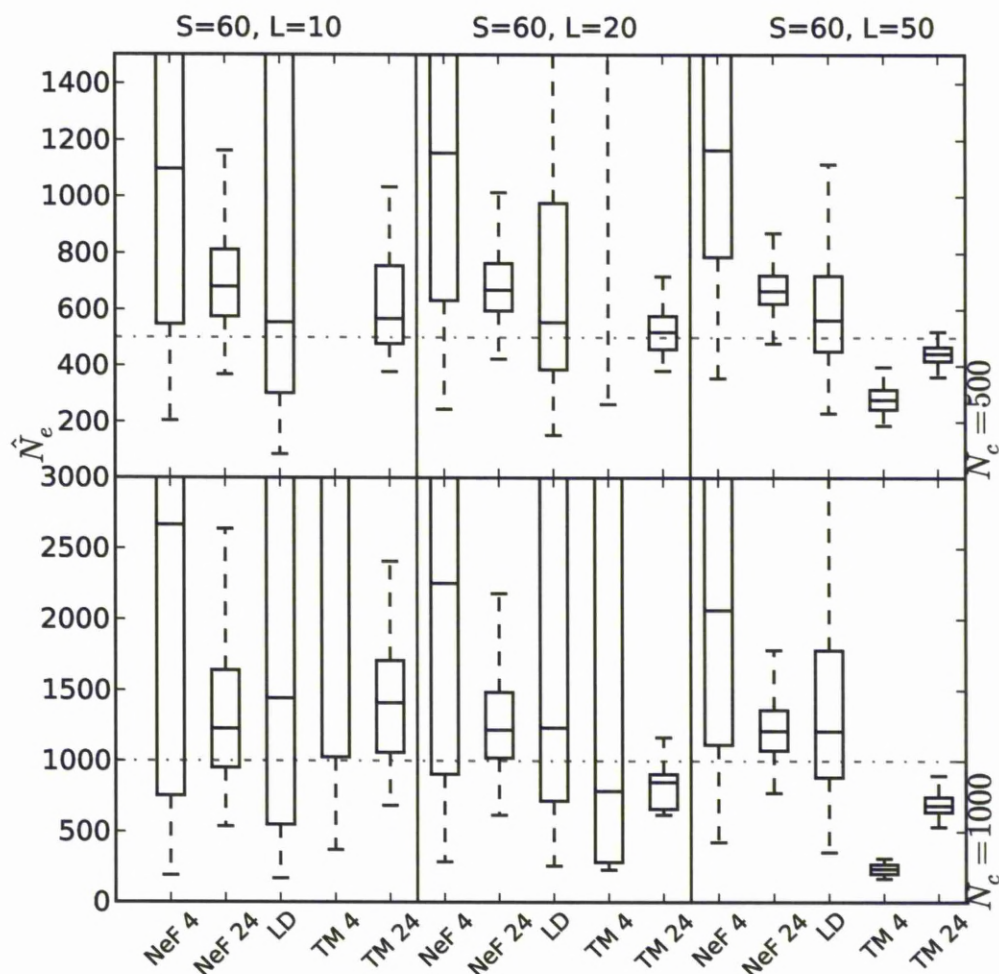


Figure 12.3: Boxplot charts for the point estimates of all three N_e estimation methods using different sampling strategies (60 individuals with 10, 20 and 50 loci) and constant N_e of 500 and 1,000. For the both temporal estimators two time spans were considered: 4 and 24 generations. The moment based temporal method is called “NeF”, and the ML version “TM”.

In terms of bias, the LD method is always less biased than any of the temporal methods for an N_e of 500. Temporal methods, for a time span of 4 generations, are in most cases less precise than LDNe and always more precise with time spans of 24, reinforcing the importance of the number of generations between samples.

Fluctuation in population size

For seasonal models, the temporal estimators and the LD method show different behaviours. The LD method is most influenced by the N_e value of the previous generation whereas the temporal methods will estimate an average between maximum and minimum N_e . This makes the LD estimator vary from generation to generation while the temporal estimators are much more stable (figure 12.4 top). While the median estimation of the LD method will approach the N_e of the previous generation, the precision of the estimations will produce overlaps of point estimate distributions (figure 12.4 bottom); the precision will be especially poor when the estimator is based on higher N_e values. Only between the extremes (N_e of 200 and 1,200) of the simulated sizes is the overlap minimal.

The LD method is able to detect a bottleneck (defined here as \hat{N}_e is below twice the post-bottleneck N_e) as early as one generation after it occurs for an N_e of 200 (figure 12.5) and 500. The temporal methods are much less sensitive, for example, nine generations after the bottleneck, the moments based estimator is still above 3 times N_e for both bottleneck scenarios. The ML estimator has lower value (i.e. is more influenced by the sample with lower N_e) than the moments based version. Again as in the seasonal model, the temporal methods average the trend spanning the sampling period while the LD method reflects the state at the sampling time.

12.4 Discussion

Our most important and novel results are that LDNe and temporal methods have different behaviour for fluctuating populations and that the re-interpretation of existing empirical studies suggests that sample sizes are often too small in order for N_e estimates to be informative. However we discuss results in order of increasing demographic complexity starting from classical temporal methods in stable populations.

Comparing temporal methods

For most scenarios tested, precision is more problematic than bias. Though our results show bias that can go up to 50% with the temporal method, precision can vary above one order of magnitude between the lower and upper quartile. Furthermore, our simulation assumption of Wright-Fisher equilibrium (i.e. $N_c = N_e$) is expected to have a relatively large bias with the standard temporal method, whereas for more realistic relationships between N_c and N_e , bias can be lower (Waples, 1989). For realistic generation spans (i.e. above 4), the sampling strategy has more influence on precision than on bias, therefore it is possible to vastly improve precision with a better sampling strategy.

Precision

With the temporal methods, small timespans between samples severely decreases precision and enlarges CIs. A clear example of this effect can be seen in Simard et al. (2000) where all estimates made with 4 months of separation (assumed conservatively to be approximately 4 generations for *A. gambiae*) include infinity in the confidence interval. In the same study, estimations with more than 3 years of interval (approximately 40 generations) do provide much tighter confidence intervals (never including infinity). The likelihood approach also provides more precision and tighter confidence intervals than the classic estimator in almost all simulated cases. Only with small time spans (4 generations) and few loci (10 or 20) is the precision of the ML estimator worse but this “advantage” is mostly theoretical as the precision of the classic estimator, while better with small time spans is still very bad, i.e., while the moments based estimator performs, in theory, better, it is still unusable. Our results suggest that the temporal methods are not useful if the time span is low and the real N_e is equal or larger than 500. Similar results were observed in Ovenden et al. (2007).

Bias

The classic moments based method produces estimates that are generally biased upwards. The strategy of binning rare alleles is not enough to eliminate the bias, though it improves results if loci with many alleles are included. Both the LD and ML methods are less biased than the moments based method. It is not clear that binning rare alleles with the ML method (as done in Dyer et al. (2009)) is a good strategy as this method is known to perform well with rare alleles (Berthier et al., 2002) and our results suggest that the estimator is normally biased downwards, thus any binning might compound the problem. The ML method is only strongly biased upwards when the sample size is too low (e.g with 35 individuals and 12 loci as in Dyer et al. (2009)). Indeed, the results in Dyer et al. (2009) where the ML estimation is higher than the classic estimate suggests that the sampling size might have been insufficient. This relationship between estimators happens with small sample size (figure 12.3) and the sample size per time point was indeed very low.

Comparing LD with temporal methods

The time span between samples is a crucial parameter in deciding which method to use. For large time spans, both temporal estimators provide more precise estimates than the single-sample LD method. From a practical perspective the temporal methods require twice the sampling effort as two time points are needed to make an estimate. In deciding which method to use, researchers should consider not only the possible time span between samples in the temporal method, but also the economic issue of using one or two time samples: much more precision can be gained from having a single sample with 120 individuals and applying the single-sample LD method instead of having two time samples of 60 individuals each.

Interpreting existing empirical studies

Our results suggest that the typical sampling policy of 60 individuals and 10 loci results in poor precision in almost all cases and severely limits biological/epidemiological interpretations of the estimates. Only with long time spans between temporal samples and where N_e is less than 500 are the estimates relatively precise. In theory increasing the number of individuals will achieve a slightly higher increase in precision than increasing the number of loci. However, given the problems with collecting some of these vector species and the advent of new high throughput genotyping platforms, a more practical suggestion in most cases may be to increase the number of loci. Using 60 individuals and 50 loci will allow for more precise estimates, if N_e is less than 1,000, though if the expected value of N_e is close to 1,000 increased spacing between samples (above 24 generations) may be required.

Next generation sequencing and SNP chips will allow the usage of thousands of SNP markers. Further research should consider the performance of estimators with thousands of bi-allelic loci with varying linkage disequilibrium. Our results also suggest that, for both methods sampling more individuals yields a bigger increase in precision than sampling more loci. This is consistent with results for the ML temporal method (Berthier et al., 2002). For the LD method most other studies (England et al., 2010; Antao et al., 2011) (but not all, see Tallmon et al. (2010)) suggest that indeed more individuals provide more information.

Higher real N_e decreased precision with all estimators, for example the published empirical temporal estimates sampled a year apart (circa 12 generations) in the vector control study of Pinto et al. (2003) either include infinity in the upper confidence interval (with a high \hat{N}_e of 1,078) in one case or even the point estimate is infinite in another case. The 24 month estimate of 1,457 has a upper confidence interval of almost one order of magnitude above (13,677). Such estimates, including infinite in the confidence intervals (and even in point estimates) make any inference of the impact of control measures unreliable at best.

Population fluctuation

When the population size is not constant the temporal and LD methods have qualitatively different behaviours. The LD method is extremely sensitive to the N_e of the previous 1 or 2 generations whereas the temporal methods “smooth” the ongoing demographic processes (figures 12.4 and 12.5). This is not a suggestion that one class of method is “better” or “worse” than the other, only that they have different applications: the LD method is better suited for early detection of bottlenecks (e.g. population reducing interventions) or to study seasonality whereas the temporal estimators provide a better picture of the “average” population size. It should be noted however that a less naïve interpretation of LD results suggests that it might not be applicable to study

seasonality as the confidence intervals of the seasonal point estimates overlap (as with higher values of N_e the precision drops). Even if the point estimate distribution could be tightened (by increasing sample sizes, especially with more individuals sampled as the LD method is substantially more sensitive to individuals than loci (Antao et al., 2011)), the confidence intervals might still overlap for a fine-grained (monthly) estimation, though detection of extremes will probably be feasible.

Any previous conclusion made about seasonality and control measures using temporal estimators is thus fraught with uncertainty. While we used the temporal ML method in Berthier et al. (2002) as it was used in a vector biology study with large N_e (Dyer et al., 2009), more recent ML temporal based methods developed to detect bottlenecks (Beaumont, 2003) should be investigated in the context of vector biology to assess their performance to detect control interventions and perhaps seasonality.

Simard et al. (2000) suggests (based on Nei and Tajima, 1981; Pollak, 1983) that, as the moments-based estimator approximates the harmonic mean of the effective population sizes, it is dominated by the smallest value, a fact also highlighted by O’Ryan et al. (1998). Our findings are consistent with this statement but we note the following: (i) While this effect is visible, the temporal estimators are still biased high from the post-bottleneck N_e , (ii) if there is an expansion, the estimator is not useful and (iii) as the time distance between the bottleneck and the second sample increases the estimator will tend to approach the contemporary value. This latter effect is possibly caused by drift.

Computational and biological assumptions

Testing the ML estimator in a wide array of scenarios is not feasible due to the high computational cost of the TM3 application (Berthier et al., 2002) (common to most likelihood approaches). While running a single instance of the application is computationally cheap, running thousands of evaluations is prohibitive. At the risk of sparking controversy, we raise the following question: if extensive testing of likelihood applications is computationally unfeasible and only limited *ad-hoc* tests are possible, can we trust the results of ML approaches?

Some simulation assumptions might require further investigation in the particular context of vector biology: while random mating might be a reasonable assumption with Anophelines, it is less clear that it is acceptable for other vector species such as *Glossina* because polyandry and sperm competition is widespread with insecta (Simmons, 2001; Tripet et al., 2001). Age structure might be of particular importance with regards to seasonality estimates with the LD method as the signal from previous generations might “smooth” the estimation curve. On the other hand, as the ratio between N_e and N_c is likely well below 1, bias with the temporal method will most probably be much lower than reported here. If a sample is mostly originating in a small breeding site where

close genetic relatedness between individuals exists, the results above might also not be applicable because the data are not independent. A practical recommendation is therefore to ensure good quality data, e.g. in the case of *A. gambiae* which breeds in small water bodies, not to take more than one individual per site.

12.5 Conclusion

Our results suggest that many existing empirical studies might require re-interpretation. It is not clear that common sampling and genotyping strategies are sufficient to reliably estimate N_e in general or to detect the impact of control measures or to estimate seasonality (fluctuation) of vector population size. While previous studies might lack sufficient sample sizes (and in some cases enough time between sampled generations), contemporary N_e estimators can reliably be used to infer population size and evaluate control interventions as long as sample size (of loci and individuals) is enough. The temporal and LD methods have different strengths and thus complementary applications: the temporal methods can provide a useful average measure of N_e over several generations of genetic drift (between temporally-separated population samples), while the LD method is sensitive enough to provide generation-to-generation estimates of N_e and detect sudden changes in population size (in populations with non-overlapping generations). Consequently the LD method can be used to assess the success of control interventions (bottlenecks) and may help infer seasonality patterns of fluctuation in effective population size. Choice of estimator type (temporal or LD) will thus be dependent on the question being asked.

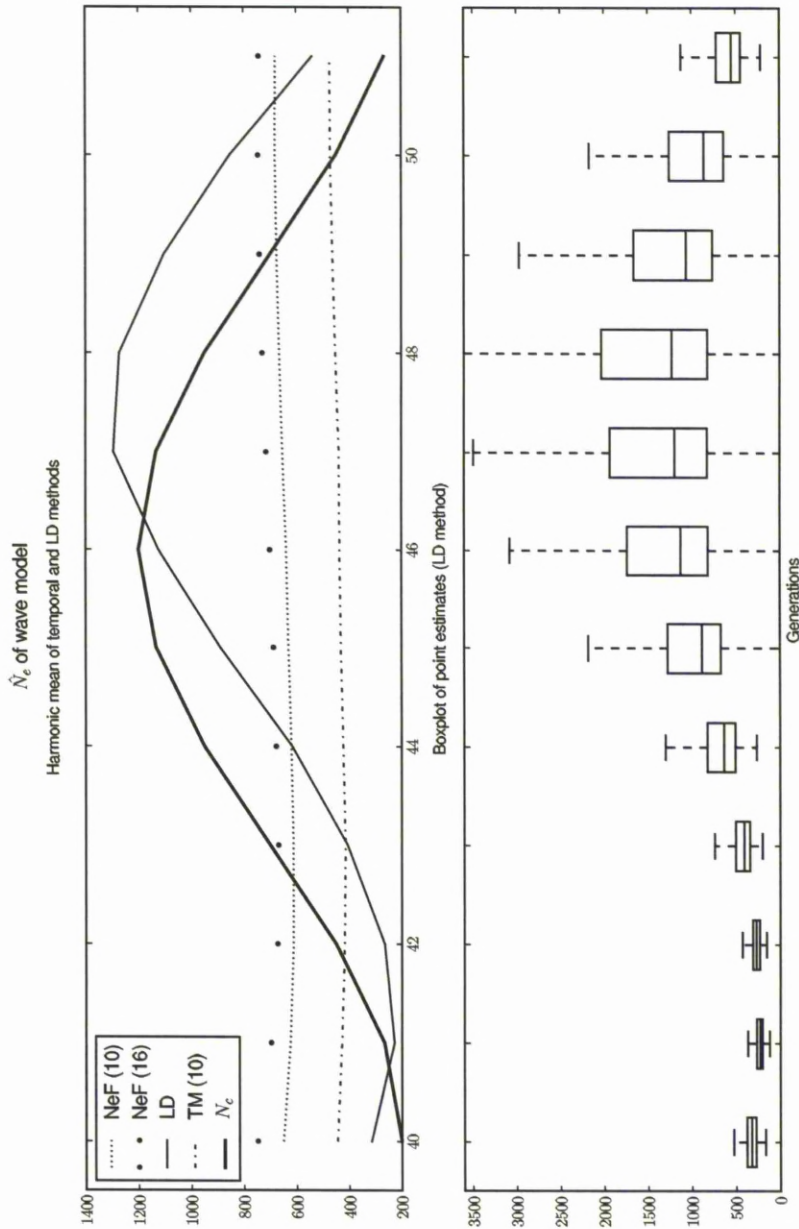


Figure 12.4: The behaviour of all estimators with a seasonal model with fluctuating population size. The top chart presents the harmonic mean of the point estimates for all estimators using 60 individuals and 50 loci. For the moments-based temporal method two reference generations were used: generation 10 with an maximum N_e of 1,200 and generation 16 with a minimum N_e of 200. The bottom chart presents the boxplot distributions of the LD estimator.

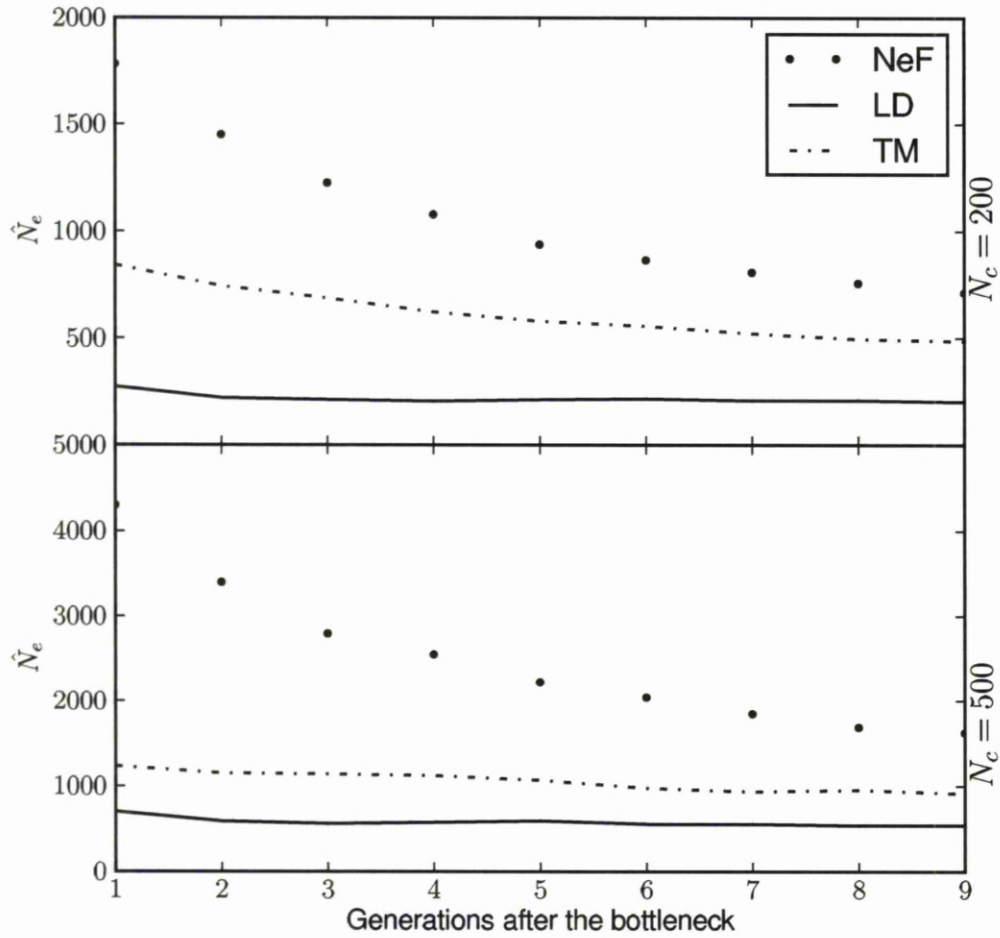


Figure 12.5: Harmonic mean of \hat{N}_E (solid line) for all methods for two bottleneck scenarios (N_e from 2,000 to 200 and from 5,000 to 500) and a sampling strategy with 60 individuals and 50 loci. Generations are counted from after the bottleneck event.

Thirteen

Interpreting estimates of effective population size in parasites and vectors

Tiago Antao, Andrés Pérez-Figueroa,
Ian M. Hastings, Martin J. Donnelly and Gordon Luikart

Abstract

Heterozygosity and effective population size are crucial parameters related to parasite and vector population viability and are increasingly used in parasitology, with potential applications to monitor and evaluate control and elimination policies. Here we argue that long-term N_e and heterozygosity estimation do not provide a reliable monitoring strategy for evaluating the efficacy of control and elimination measures and that current research might be incorrectly using such estimators. We suggest that estimation of the contemporary N_e might be used to evaluate control and elimination measures as long as the genetic and population samples are large. We provide suggestions on how to obtain precise and unbiased estimates of both the long-term and the contemporary effective population size. We note that studies of species such as *Plasmodium falciparum* cannot employ available Linkage Disequilibrium based estimators of N_e due to its mating system which includes selfing.

13.1 Introduction

Effective population size (N_e) determines the rate of neutral genetic change in a population caused by genetic drift (Charlesworth, 2009), consequently it determines the relative efficiency of natural selection in the face of drift. N_e has many biological applications, being a fundamental factor in determining population viability as greater N_e allows more genetic variability which is paramount for species survival and adaptation. N_e is also commonly estimated in population genetic studies of *Plasmodium falciparum* at least since Anderson et al. (2000a) and of *Anopheles gambiae* starting from Lehmann et al. (1998). N_e estimation has, also in parasitology, many potential applications from

simply assessing the genetic variability and size of parasite populations (Anderson et al., 2000a; Iwagami et al., 2009; Susomboon et al., 2008) or vectors (Lehmann et al., 1998), to assessing the impact of control interventions on parasites (Gatei et al., 2010) and vectors (Pinto et al., 2002, 2003; Wondji et al., 2005) and monitoring vector seasonality (Simard et al., 2000) (seasonality fluctuations in population size).

Estimating N_e mainly uses two different approaches (for reviews see Leberg (2005); Wang (2001)):

1. Long-term estimation of N_e is based on the observation that at mutation-drift equilibrium the amount of genetic variation is stable and dependent on mutation rate. This is applicable to DNA sequences (used with *P. falciparum* in Hughes and Verra, 2001) or microsatellites where the equilibrium heterozygosity can be approximated (assuming mutation-drift equilibrium) and that mutation rate can be estimated.
2. Short-term (contemporary) N_e estimators can be useful if population effective sizes are small, e.g. $N_e < 500$, but for many parasites and vectors N_e is often too large. Contemporary N_e is normally estimated from observed changes in allele frequencies between generations (Krimbas and Tsakas, 1971; Pollak, 1983), from linkage disequilibrium patterns (Hill, 1981; Waples and Gaggiotti, 2006) or heterozygote excess (Robertson, 1965; Luikart et al., 1998; Balloux, 2004). The estimators of the contemporary N_e measure either the variance or inbreeding N_e for details see (Schwartz et al., 1999; Leberg, 2005).

Estimations of N_e in *P. falciparum* have been computed using long-term estimation approaches presumably because N_e is relatively large and contemporary estimators of N_e are not precise with high N_e values (Waples, 1989; Waples and Do, 2010). *P. falciparum* effective population size and thus its population genetic diversity increases with transmission intensity with estimations varying between 1,000 for low transmission areas in South America and 20,000 in high transmission areas like Central Africa (Anderson et al., 2000a). Estimation is commonly done using an heterozygosity based estimator with microsatellites (see e.g. Anderson et al., 2000a; Iwagami et al., 2009; Susomboon et al., 2008), therefore the proprieties of heterozygosity estimators have direct consequences on this kind of long-term N_e estimation. As far as we know, estimates of contemporary N_e have never been reported for *P. falciparum*, although in areas of low transmission (due to expected low N_e) it might be possible to estimate contemporary N_e with precision, especially if the sampling strategy includes many individuals and loci.

For vector species both contemporary and long-term estimators have been routinely reported (e.g., Simard et al., 2000; Pinto et al., 2003; Dyer et al., 2009). Indeed one of the most commonly used contemporary N_e estimator (Krimbas and Tsakas, 1971) was developed to study the impact of insecticide resistance. This method requires two

temporal samples and computes the variance in allele frequencies over time (as the variance is directly correlated with the effective size and decreases with N_e).

Here we provide arguments supporting the following conclusions: 1) We cannot use long-term N_e estimators and heterozygosity, to monitor the effectiveness of control and elimination interventions, 2) Heterozygosity based estimators N_e are not informative of contemporary demographic processes, 3) contemporary estimators might be useful if the real N_e is relatively low (<1000) and 4) linkage disequilibrium estimation is probably inappropriate for species like *P. falciparum* even with low N_e . We also provide guidelines on how to more precisely estimate and interpret N_e .

13.2 Problems with heterozygosity based N_e estimation

Three problems with heterozygosity based N_e estimation exist:

- Different N_e concepts based on heterozygosity exist including long-term and contemporary N_e (as well as local and global N_e), but authors are often unclear about which concept is being used.
- Misuse occurs such as the use of the long-term N_e estimator to study the contemporary N_e .
- Most importantly, low sensitivity is problematic for both the long-term and short-term N_e estimators based on heterozygosity. This is especially serious when N_e is large, precisely the typical scenario of parasitology studies.

For example, in a bottleneck there will be heterozygosity loss at a rate of $\frac{1}{2N_e}$ per generation, where N_e is the effective population size after the bottleneck event. Assuming the data presented by Anderson et al. (2000a) that a high-transmission zone has a N_e of around 18,000 (H_e of 0.8) and a low-transmission zone has an N_e of approximately 1,400 (H_e of 0.4) then a control intervention that would cause a reduction from 18,000 to 1,400 would take 1,300 generations (more than 260 years in the case of *P. falciparum*) just to approximate an H_e of 0.50 (N_e of $\approx 2,500$). Almost 2,000 generations would be needed to approximate a H_e of 0.40. To put this in perspective, if the usage of Chloroquine (introduced around 1947) were to impose a continuous bottleneck of the intensity described above, we would get a reasonably precise estimation of the bottleneck intensity by sampling around the year 2350. This short rate of changes illustrates that there is essentially no relationship between changes in heterozygosity (and any heterozygosity based N_e estimator, short- or long-term) and contemporary demographic processes. More detailed illustrative examples are presented in appendix C.

These results have direct application on the interpretation of the results presented in several papers. For instance Gatei et al. (2010), compared parasite heterozygosity before and after the introduction of ITNs; the results showed that there was no significant

difference in expected heterozygosity, which led the authors to conclude that the population maintained “overall stability in genetic diversity”. (Gatei et al., 2010) computed the expected heterozygosity using eight “neutral” microsatellites before the introduction of ITNs and compared the result with a sampling 5 (approximately 25 *P. falciparum* generations) years after. This was done in western Kenya, a high-transmission area. ITN use causes a substantial reduction in malaria cases and human deaths (World Health Organization, 2008), thus potentially reducing the size of the parasite population. The authors actually report an increase in heterozygosity from 0.75 to 0.79. With regards to heterozygosity and the prevalence of mixed infections the authors state:

The stable overall genetic diversity after dramatic reduction in transmission intensity observed in the current study was unexpected by the initial prediction. The counter-intuitive results suggest that other factors may be involved in offsetting the effect of transmission reduction on parasite genetic diversity and/or stabilisation of the overall genetic diversity of *P. falciparum* parasite.

It seems that that the results are neither “counter-intuitive” nor “unexpected”: expected heterozygosity is a slow moving indicator especially in high-transmission areas even when efficient control measures (i.e. imposing strong bottlenecks) are in place. What Gatei et al. (2010) is observing is expected “artifact noise.” In this study, conclusions are probably over-pessimistic: the ITN intervention might be having a impact on parasite diversity. We chosen this study just as an example among others where similar interpretations were made. For instance similar arguments could be raised for a study (Pinto et al., 2002) on *A. gambiae* about the impact of indoor spraying with DDT or several other vector studies which make contemporary inferences using the N_e estimator based on heterozygosity. The use of the DNA based long-term estimator in Hughes and Verra (2001) to infer about recent population bottlenecks probably suffers from similar problems.

Long-term N_e estimation is also influenced by the sample size (i.e. the number of individuals and loci sampled) as it influences precision and bias, but this problem (further detailed in the appendix C) is of considerable less importance than problems caused by misinterpretations of long-term N_e estimators for contemporary events. Loci under selection will also bias the estimator and, for instance, in Susomboon et al. (2008) it is suggested that 3 of the 12 microsatellite loci used to estimate N_e had strong genetic differentiation between samples taken in severe and uncomplicated malaria patients, though no formal test for selection was conducted. To our knowledge, no studies with long-term estimation in parasites or vectors have included tests for selection. Tests to detect selection should always be performed because loci documented as neutral in the past can be under selection in current or future studies (e.g. different drugs or insecticides).

13.3 Contemporary estimation of N_e

Temporal estimators

Contemporary estimation of N_e is mostly used for vectors (see e.g. Dyer et al. (2009); Pinto et al. (2003); Simard et al. (2000); Lehmann et al. (1998)). To our knowledge no studies with *P. falciparum* have used these estimators. Most studies of contemporary N_e use temporal estimators which require two samples over time. Precision of such N_e estimators is poor with large N_e , small temporal spacing between samples, and a low number of individuals or loci sampled. Figure 12.1 (from chapter 12) shows the impact of these factors on estimator precision and they are further discussed in Box 1. Very high N_e and few generations between samples are fundamental factors causing low precision. Chapter 12 shows that commonly used temporal based methods might not be appropriate to detect bottlenecks (normally resulting from control interventions) or seasonality patterns. Temporal based methods are more appropriate to detect averages over a period of time. For early detection of interventions, the usage of methods based on linkage disequilibrium (LD) was recommended instead (Antao et al., 2011).

Using linkage disequilibrium to estimate N_e

P. falciparum biology is known to diverge from standard population genetic models. For instance, selfing is common especially in low transmission areas (Arnot, 1998). While *P. falciparum* is not clonal, selfing will impact LD (de Meeûs and Balloux, 2004) as associations between loci are maintained for several generations in clonal populations. This has two important consequences: contemporary estimators of N_e based on LD (Hill, 1981; Waples and Gaggiotti, 2006; Waples and Do, 2008) will probably produce erroneous results as LD patterns are maintained over time. Also, any study in the change of multi-locus LD will probably also be slow moving. Therefore it is probable that LD is also of little use to evaluate the impact of malaria control measures (though, for other fully sexual, diploidy parasites or vectors, LD might be useful).

13.4 Conclusions and guidelines

Misinterpretation of N_e and heterozygosity estimation results can occur in parasitology studies such as assessments of control and elimination measures. Heterozygosity is widely known to be an insensitive indicator, thus is not appropriate to detect recent, sudden changes in population size. The larger the N_e value, the slower the change in heterozygosity. We illustrated, in the appendix, that poor performance expectations hold for typical effective population sizes of *P. falciparum*. Even for extreme bottlenecks the N_e estimated after 100 generations (i.e. around 20 years, assuming 5 *P. falciparum* generations per year) is still closer to the original value than to the post-bottleneck value.

Indeed for areas that have had continuous high-transmission, it is not clear that estimates of N_e based on heterozygosity can detect any effect from the introduction of treatment drugs like Chloroquine decades ago. While theoretical predictions and simulation studies mainly done in conservation genetics would easily predict this estimator to be insensitive, it is staggering that some interventions in the distant past, especially in high transmission areas, cannot be detected even today using this estimator.

We provide suggestions to help future studies of effective population size, heterozygosity and LD for *P. falciparum* malaria or insect vectors:

1. Long-term N_e estimators should not be used to infer contemporary demographic processes. Little can be inferred from long-term N_e estimators as to the impact of control and elimination measures. The same is valid for heterozygosity measurements.
2. Estimators of contemporary N_e cannot be readily applied to *P. falciparum* malaria or other organisms if N_e is above approximately 2,000 as precision decreases with increasing N_e value.
3. The number of loci and individuals sampled should be carefully considered, especially when using contemporary N_e estimators (Details can be seen in e.g. Antao et al. (2011); Tallmon et al. (2010); Berthier et al. (2002); England et al. (2010); Waples and Do (2010)). Most studies suggest that sampling more individuals is more beneficial than sampling more loci (see e.g. Antao et al., 2011; England et al., 2010). If the number of individuals sampled cannot be increased then sampling more loci (feasible with the advent of next generation sequencing) can increase estimation precision of N_e . Further research is needed to understand if genotyping thousands of SNPs (current studies use microsatellites) can increase precision, especially of contemporary estimators.
4. With temporal based estimators of contemporary N_e , generations between samples should be as large as possible as precision increases as the generation spacing increases.
5. Loci should be tested for selection as selection can bias all N_e estimators. While testing for selection is common between populations, in the specific case of parasites and vectors, selection over time for the same population is also relevant due to the impact of human interventions (e.g., drug deployment and subsequent selection for resistance). Many methods developed for spacial selection detection can be easily used to detect temporal selection signatures.
6. In the specific case of *P. falciparum* and due to reproduction with selfing, inferences based on LD (including contemporary N_e estimators based on LD) should be used

with great care, most probably avoided. Further research is necessary to evaluate these estimators for use with *P. falciparum* and parasites with similar genetics.

7. The appropriate contemporary estimator should be correctly chosen for different research problems. The LD estimator is probably better suited for detecting the impact of interventions and seasonal fluctuations, whereas temporal based approaches can be used to estimate an average N_e over multiple generations. Temporal estimators exist for detecting population fluctuations (see e.g. Beaumont, 2003), but have never been used or evaluated in parasite or vector contexts.
8. Finally, both parasite or vector populations are not closed and migration is normally an important factor, especially considering modern human mobility. Migration will bring increase genetic variability in local populations and thus increase short-term N_e .

While it would be important to have an accurate and precise estimate of contemporary effective population size for *P. falciparum*, its vectors and other parasites, especially in the context of control and elimination measures, it is not clear that that objective is always feasible: methods explicitly targeted for estimating contemporary N_e are only applicable for small populations or will need large amounts of data. Furthermore, heterozygosity based methods cannot reliably estimate contemporary N_e , unless it is extremely small. Some of these limitations are widely known, but unfortunately it is not clear that they have been understood by the parasitology community. We recommend caution in interpreting N_e estimates and suggest that other strategies should be used to complement assessments of impacts of control and elimination measures on parasite and vector genetic diversity.

Box 1: Factors affecting the precision of contemporary N_e estimators

The precision of the most widely used temporal-based method for contemporary estimation of N_e is well characterised by its coefficient of variation, which is a measure of dispersion:

$$CV(\hat{N}_e) \approx \sqrt{\frac{2}{n}} \left[1 + \frac{2N_e}{tS} \right] \quad (13.1)$$

In the context of this work, more important than any mathematical details, it is the intuitive consequences that are important:

Precision increases with n (the number of independent alleles sampled), S (the number of individuals sampled) and t (the spacing, in generations, between samples).

Interestingly precision to estimate N_e decreases with the real N_e value. This means that the higher the N_e the lower the precision obtained (all other variables being equal). There is a relationship between the real value and the precision of the estimator.

One of the most widely used linkage disequilibrium estimators is implemented in LDNe software application. Its coefficient of variation is also known and is:

$$CV_{LD}(\hat{N}_e) \approx \sqrt{\frac{2}{n'}} \left[1 + \frac{3N_e}{S} \right] \quad (13.2)$$

Where n' is a slight formal variation of n above, but from a qualitative perspective represents the same concept (roughly the number of alleles).

The factors affecting precision of this LD estimator are similar to the temporal method (except the temporal spacing between samples, as this LD method uses only a single sample).

There is ample literature about the influence of these parameters considering many demographic situations, but most of these studies are normally geared for conservation genetics, thus the reader is cautioned that the N_e s studied are normally lower than the typical values in parasitology settings, thus reported recommendations will be insufficient due to decreased precision. Researchers are thus strongly recommended to verify if the conditions of the literature consulted are applicable to their species and research question, and in the scope of parasitology.

Part IV

Discussion and conclusion

Fourteen

Discussion and conclusion

This thesis follows, very consciously, a strongly empiricist approach. This might seem strange at first: in effect this is a theoretical thesis with apparently no empirical data supporting it. As we will see, there is absolutely no contradiction in a theoretical work being empirical in approach. On the other hand, as we will also see, most of this thesis is indeed supported by data. But first we will address the empiricist slant of this work.

The vast majority of existing theoretical work in parasitology is mostly concerned with the *future*: from predicting the impact of new vaccines (Smith et al., 2006), to models of economic cost (Tediosi et al., 2006) or proposals for malaria elimination (Maude et al., 2009) among many others. With a single exception (chapter 4) this work is concerned with data interpretation and understanding the *past* and the *present*. Even the most theoretical part, dedicated to population genetics models of drug resistant *P. falciparum* is mostly empirically driven: Chapter 3 tries to understand the cause of stabilization of resistance at intermediate frequencies and provides guidelines to better understand linkage disequilibrium. The same pattern occurs in the part dedicated to selection detection: Chapter 9 uses empirical data with a known demographic and selection history (humans) to assess selection detection methods based on F_{ST} and chapter 10 studies the performance of F_{ST} using simulated data. Chapters 6, 7 and 8 are software applications to analyze empirical data. Exactly the same pattern can be observed in the final part, dedicated to N_e estimation: all chapters are concerned with suggesting realistic sample sizes to infer N_e , furthermore the applicability of several estimators is also assessed for different demographies (e.g. bottlenecks caused by control policies or seasonality) or parasite biology (e.g. the impact of selfing on linkage disequilibrium).

It should be clear by now that it is possible to be theoretical and empirically oriented at the same time. It is actually puzzling (and worrying) that theory and practice seem to be so disconnected within the parasitology community:

First, there is no such thing as “a-theoretical” empiricism: a good example of this is the (arguably wrong) expectations about positive linkage disequilibrium between resistance genes. There is clearly a (often unconscious) theoretical model of linkage being used (namely that there should be a positive association between resistance loci). And

one has to wonder how many studies of linkage between, e.g. *dhfr* and *dhps*, are done where equilibrium is found and which are subsequently not reported due to fear of being seen as a “wrong” result.

Secondly, theoretical work seems to concentrate on predicting the future: there seems to exist an implicit view that the past is somehow perfectly understood, almost as if existing knowledge does not require revision. Also, there is clear pressure to concentrate on the future, for instance several funding bodies seem to desire, at any cost (especially a cost of realism), predictions and strategies to eliminate and eradicate malaria. In this context reflections on the past are of little interest. But understanding the past is important: it is nothing more than a fallacy that we have full (or enough) understanding of it. Several examples of the need to revisit the past and re-interpret data and associated theories were presented here, for instance we noted that exiting interpretations of heterozygosity and effective population size might be too pessimistic when assessing the impact of control measures on parasite and vector genetic diversity.

It should also be clear by now that empirical data is used: Obviously we use human data from the HapMap project in chapter 9, but for most other chapters we made sure that our models qualitatively followed field observations and great care was put on assuring that our models and computational approaches provided results that were consistent with observations. Our approach to “fitting” was qualitative, not quantitative: there is no presumption that our simple models will track reality with quantitative precision. Interestingly there is a question if complex models can do any better, but that is a question that we did not address here. Sometimes we used simulated data for analysis, this was always done with an empirical perspective: for instance to try to understand the correct sample sizes to precisely estimate N_e . Simulated data is fundamental when there is no empirical data available, such can happen in many cases and especially with human diseases like malaria where experimental design and data collection are strongly limited by ethical issues.

Being a theoretical thesis with an empirical approach makes for inherently modest work: Here there are no grand plans on how to eliminate and eradicate malaria. This thesis is mostly comprised of suggestions for researchers that do data analysis. Therefore, while this work is theoretical in nature, it is intended with an empirical readership in mind. It is, in a way, an attempt to approximate the two communities (theoretical and empirical).

After this more philosophical discussion (a gentle reminder that this is a thesis submitted for a *Philosophical* doctor degree) I provide a conclusion for each part of this thesis followed by some final remarks.

14.1 Population genetics models of drug resistant *P. falciparum*

P. falciparum biology has several differences from “standard” population genetics models. Three differences are worth noting here: having both an haploid and diploid phase; large population size inside humans (up to 10^{12}) contrasting with low size in mosquito blood meals (sometimes below 10) and highly inbred mating (due to limited mating options during the sexual phase inside the mosquito). OgaraK was thus developed to allow the simulation of *P. falciparum* population genetics in the presence of drug resistance. This is a novel simulator but very limited in features (only simulates resistance genes, with only two alleles per gene, etc...), this became clear in the chapter on the importance of F_{ST} in *P. falciparum* studies, when there was a need to use a different simulator to calculate neutral loci. There is evidently room for improvement and future work might include the development of a more general simulator of *P. falciparum* population genetics. Even in its current version, some features are still missing: for instance it is not possible to simulate other interactions than full epistasis when simulating combination therapies. OgaraK is fully open-source allowing for easy repetition of the results presented here (and the creation of many other novel simulations). I wonder why such openness is not compulsory. If researchers are required to fully report all mathematical formulae, why not extend such approach to software applications developed to conduct research?

Chapter 3 extended the model of Hastings (2006) to support epistasis and multiple drug deployments. It is the basic theoretical work of this part and sets the theoretical ground to all other research. There are three main conclusions that are worth revisiting:

1. The impact of epistasis on linkage disequilibrium. Our results provides a sound theoretical base suggesting that any kind of linkage disequilibrium (in either magnitude or signal) are plausible. Most notably, researchers should not be reluctant to report linkage equilibrium. The magnitude and signal of linkage provide important information as to the interactions among genes involved in drug resistance.
2. This chapter also provides a novel justification for the stabilization of resistance at intermediate frequencies. As chapter 10 suggests while stabilization of allele frequencies might occur, the resistant phenotype might still, in essence, fixate. This suggests that, while epistasis is surely part of the explanation for stabilization at intermediate frequencies, further research is still needed to understand all the causes of stabilization.
3. The mis-interpretation of changes in prevalence of resistance (normally caused by control interventions). If MOI is reduced, then any observed decrease in prevalence of resistance might be explained not by a reduction of frequency of resistance, but

simply by MOI. Analysis of field data where transmission has changed should carefully consider this.

As most of the chapters of this thesis were accepted for publication before submission, content, and especially format was affected by restrictions imposed by accepting journals. For instance, chapter structure varies considerably and content is sometimes repeated in different chapters. In terms of content, only chapter 3 requires further discussion: Firstly, the original title was (the arguably more modest) “The influence of selection heterogeneity on the spread of drug resistant malaria”, this title puts more emphasis on the evolutionary approach of the model, whereas the final title, “Environmental, pharmacological and genetic influences on the spread of drug resistant malaria”, is more broad in scope. Secondly, and most importantly, the third paragraph “Mathematical models play an important role...” engages in what this author sees as proselytising about the role of mathematical models, I would clearly have preferred a more “agnostic” approach: if mathematical models are important, that is for the reader to make her own judgement. But, as a rule, reviewer and editor comments improved the quality of the manuscripts in substantial ways.

A final, important, note about this chapter: In hindsight, the fundamental messages (for the intended target audience, empirical researchers) are written in the middle of many theoretical formulations and discussion. It is not clear that such format is the best to convey such important information to empirical researchers. A more compact and readable formulation would probably have been desirable.

Chapter 4 is the only chapter that is, in its nature, predictive. Due to serious problems with multi-drug resistance in other infectious diseases (e.g., MRSA or tuberculosis), it was deemed important to analyse the impact of recent proposals regarding drug distribution policy which make available several first-line treatments simultaneously. There will probably be no future attempt to publish this chapter.

This part finalises with a small opinion letter (chapter 5) reflecting on the importance of evolutionary approaches in parasitology; most importantly we try to remind the policy makers that elimination attempts, if unsuccessful might drive resistance dangerously higher.

14.2 F_{ST} selection detection and discovering genes involved in drug resistance

LOSITAN was developed to study the performance of a widely used F_{ST} -outlier method (Beaumont and Nichols, 1996) to identify loci under selection. Most unfortunately there was never real *P. falciparum* data to be studied (discussed below). Interestingly the application has been widely used to study selection accross many taxa. It is nonetheless gratifying to know that is was used elsewhere (Vinayak et al., 2010) to study the spread

of resistance to Sulfadoxine in *P. falciparum*. LOSITAN is mostly an easy to use application based on the FDIST program, but its features are much more than a simple interface: it includes a multi-test implementation, removal of potentially selected loci when calculating neutral F_{ST} and correct approximation of neutral F_{ST} . Many selection detection methods (or much biological software, in general) are difficult to use, and consequently prone to being mis-used. Developers of biological applications tend to assume that the users will be fully knowledgeable about the underlying parameters of biological models, here we follow a different approach: we try to help the user avoid making parametrization mistakes.

Mcheza is proof that the work conducted here was prone to some fluidity. This application was never planned, but many researchers using LOSITAN requested a version for dominant markers and therefore we decided to develop and provide it.

interPopula is a library to process HapMap data in Python. It was developed with the sole purpose of supporting the evaluation of F_{ST} with HapMap data. The most serious drawback of this library comes from the fact that the HapMap project is constantly changing the directory structure of the HapMap repository, making it difficult to maintain the library compatible with the repository format.

Chapter 9 evaluates F_{ST} using the HapMap SNP dataset. Our approach is novel as we use empirical data to evaluate performance, whereas all known studies use simulated data. The strategy is simple: There is relatively much more data available for humans (both in terms of dataset size and information about demography and selection) so we can compare the results using a more limited methodology (either in data size or knowledge about demography and selection): If the more limited methodology is able to approach the complete information available, then the more limited methodology is reliable. Our results suggest that, while F_{ST} is “noisy” over the genome, it is far from being randomly distributed. Current sampling strategies systematically under-sample the number SNPs per gene necessary to reliably detect selection and probably fail to detect many candidate loci for directional selection. Most disturbingly, some next generation sequencing technologies might still not have enough markers per gene in order to reliably detect loci under selection.

There is an ongoing debate on the usefulness of F_{ST} -outlier approaches to detect selection, this part is contribution to that debate, which is clearly far from over.

14.3 Estimating effective population size and assessing the success of control and elimination measures

To study the performance of effective population size estimators in parasitology settings we start, in chapter 11, by comparing two short term N_e estimators for bottleneck

detection, but still in low N_e (e.g. conservation) scenarios. There are several reasons for this approach:

1. Short term N_e estimators are known to have decreased precision as real N_e increases, therefore low N_e is a starting point before higher N_e values are considered: if estimators were to fail in low N_e scenarios, there would be no point in proceeding to more computationally expensive simulations of higher N_e .
2. Our evaluation was novel as it evaluated bottlenecked demographies: precisely the type of signal that we want to detect in any control and elimination intervention.
3. Most importantly we wanted to attract reviewers and readership with a conservation background. We will argue below that such audience is probably more capable of providing insightful feedback and constructive criticism than parasitologists.

Chapter 12 is another example of a certain degree of fluidity in the development of this work: Studying N_e in disease vector scenarios was never planned but, as a review of the broad parasitology literature on effective population size made clear, it was important to understand the performance of N_e estimators in the context of vector biology. Our demographic model designed to approach seasonality (i.e., the impact of dry- and wet-season on population size) clearly demonstrated that temporal and LD based N_e estimators have qualitatively different interpretations: LD estimators are very short-term as they estimate the N_e of the previous generation whereas temporal estimators tend to provide an average over a period of time. We also showed that existing studies of N_e have an excessively small sample size and that, in some cases the interval between generations used with temporal estimators is too short. Nonetheless, as long as sampling sizes and temporal spacing between generations are both increased and care is taken with the interpretation of results, N_e can be a useful tool to understand the impact of vector control measures.

Finally, chapter 13 describes a broad analysis of the usage of effective population size estimators in parasitology, especially long-term estimation. It is puzzling, not to say shocking, that many basic interpretation mistakes are routinely being made in the interpretation of genetic data. The most obvious is, of course, the use of heterozygosity (and heterozygosity based N_e estimation) to make conclusions about potential recent bottlenecks. Basic population genetics postulates that heterozygosity is “slow moving”, and slower with higher population sizes. It has been shown that even for low N_e scenarios heterozygosity cannot be used for early detection of bottlenecks. It is worrying that such results, seen as trivial by most conservation geneticists are not absorbed by the parasitology community. Given the existing number of manuscripts making such erroneous interpretations one can infer that the problem is pervasive: it affects authors, reviewers and editors to a reasonable extent.

It is clear that many biological assumptions made in standard population genetics models are not applicable to *P. falciparum*. There is clearly much work to be done in both developing new models targetting malaria and understanding the implications (and possible mistakes) of using standard assumptions with *P. falciparum*. For instance, and in the scope of this part: Quantifying the error of LD-based N_e estimation imposed by selfing in particular and inbreeding of *P. falciparum*.

14.4 Final remarks

The direction taken by this work was mostly decided after the first year where I started to pursue a empiricist-based strategy, as opposed to a predictive approach (where most of the first year work was done). It was clear from the onset that having real datasets to work with would be a difficult objective to attain: several possible sources of real data were considered but it became obvious that access to them would not be possible in time to complete this thesis. The responsibility of the path followed relies solely on me as I was fully aware of the risks involved. Despite all this, one year lost and no *P. falciparum* datasets to work with, I would, without a shadow of a doubt, repeat the same approach: for all time lost and obstacles are easily compensated by doing work in which I believed and which was, from my point of view, valuable and worthwhile. Such change in strategy and scope requires obviously a very special kind – a rare kind – of supervisor: one that understands the value of intelectual freedom and is always there to support difficult, but necessary, decisions.

One of the aims of this work is to show that any supposed barriers between theory and practice are mostly fabricated. It is possible, and in most cases desirable, to theorize about subjects which are important to field researchers.



Bibliography

- Abdel-Muhsin, A, Mackinnon, M, Awadalla, P, et al. Local differentiation in *Plasmodium falciparum* drug resistance genes in Sudan. *Parasitology*, 126(05):391–400, 2003. ISSN 1469-8161.
- Akey, J, Zhang, G, Zhang, K, et al. Interrogating a High-Density SNP Map for Signatures of Natural Selection. *Genome Research*, 12:1805–1814, 2002. 10.1101/gr.631202.
- Allendorf, F and Seeb, L. Concordance of genetic divergence among sockeye salmon populations at allozyme, nuclear DNA, and mitochondrial DNA markers. *Evolution*, 54(2):640–651, 2000.
- Altshuler, D, Gibbs, R, Peltonen, L, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467:52–58, 2010.
- Amato, R, Pinelli, M, Monticelli, A, et al. Genome-Wide Scan for Signatures of Human Population Differentiation and Their Relationship with Natural Selection, Functional Pathways and Diseases. *PLoS ONE*, 4(11):e7927, 2009.
- Anderson, T. Mapping drug resistance genes in plasmodium falciparum by genomewide association. *Current Drug Targets-Infectious Disorders*, 4(1):65–78, 2004.
- Anderson, T and Roper, C. The origins and spread of antimalarial drug resistance: lessons for policy makers. *Acta tropica*, 94(3):269–280, 2005. ISSN 0001-706X.
- Anderson, TJ, Haubold, B, Williams, JT, et al. Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Mol Biol Evol*, 17(10):1467–1482, 2000a.
- Anderson, TJ, Su, XZ, Roddam, A, et al. Complex mutations in a high proportion of microsatellite loci from the protozoan parasite plasmodium falciparum. *Mol Ecol*, 9(10):1599–1608, 2000b.
- Antao, T. interPopula: a Python API to access the HapMap Project dataset. *BMC bioinformatics*, 11(Suppl 12):S10, 2010.

- Antao, T. Evolutionary parasitology applied to control and elimination policies. *Trends Parasitol*, pages –, 2011. ISSN 1471-4922.
- Antao, T and Beaumont, M. Mcheza: A workbench to detect selection using dominant markers. *Bioinformatics*, 2011.
- Antao, T and Hastings, I. ogoraK: A population genetics simulator for malaria. *Bioinformatics*, 2011a. ISSN 1367-4803.
- Antao, T and Hastings, IM. Environmental, pharmacological and genetic influences on the spread of drug-resistant malaria. *Proceedings of the Royal Society B: Biological Sciences*, 278(1712):1705–1712, 2011b.
- Antao, T, Lopes, A, Lopes, R, et al. LOSITAN: a workbench to detect molecular adaptation based on a F_{ST} -outlier method. *BMC bioinformatics*, 9(1):323, 2008. ISSN 1471-2105.
- Antao, T, Perez-Figueroa, A, and Luikart, G. Early detection of population declines: high power of genetic monitoring using effective population size estimators. *Evolutionary Applications*, 4(1):144–154, 2011.
- Arnot, D. Unstable malaria in Sudan: the influence of the dry season. Clone multiplicity of *Plasmodium falciparum* infections in individuals exposed to variable levels of disease transmission. *Trans R Soc Trop Med Hyg*, 92(6):580–585, 1998.
- Atwal, GS, Kirchhoff, T, Bond, EE, et al. Altered tumor formation and evolutionary selection of genetic variants in the human MDM4 oncogene. *Proceedings of the National Academy of Sciences*, 106(25):10236–10241, 2009.
- Babiker, H, Hastings, I, and Swedberg, G. Impaired fitness of drug-resistant malaria parasites: evidence and implication on drug-deployment policies. *Expert review of anti-infective therapy*, 7(5):581–593, 2009.
- Babiker, H, Satti, G, Ferguson, H, et al. Drug resistant *Plasmodium falciparum* in an area of seasonal transmission. *Acta tropica*, 94(3):260–268, 2005. ISSN 0001-706X.
- Balloux, F. Heterozygote excess in small populations and the heterozygote-excess effective population size. *Evolution*, 58(9):1891–1900, 2004.
- Bate, R, Coticelli, P, Tren, R, et al. Antimalarial drug quality in the most severely malarious parts of Africa—a six country study. *PLoS One*, 3(5):e2132, 2008.
- Beaumont, M. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139, 2003.

- Beaumont, M. Selection and sticklebacks. *Molecular Ecology*, 17(15):3425–3427, 2008. ISSN 1365-294X.
- Beaumont, M and Balding, D. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, 13(4):969–980, 2004.
- Beaumont, M and Nichols, R. Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society B*, 363:1619–1626, 1996.
- Beaumont, MA. Adaptation and speciation: what can F_{st} tell us? *Trends Ecol Evol*, 20(8):435–440, 2005.
- Benjamini, Y and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 57(1):289–300, 1995. ISSN 0035-9246.
- Berthier, P, Beaumont, MA, Cornuet, JM, et al. Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. *Genetics*, 160(2):741–751, 2002.
- Biller, H and Grand, R. Lactose intolerance. *Annual Review of Medicine*, 41(1):141–148, 1990.
- Boni, MF, Smith, DL, and Laxminarayan, R. Benefits of using multiple first-line therapies against malaria. *Proc Natl Acad Sci U S A*, 105(37):14216–14221, 2008.
- Bourret, V, O'Reilly, P, Carr, J, et al. Temporal change in genetic integrity suggests loss of local adaptation in a wild Atlantic salmon (*Salmo salar*) population following introgression by farmed escapees. *Heredity*, 2011.
- Caballero, A, Quesada, H, and Rolan-Alvarez, E. Impact of amplified fragment length polymorphism size homoplasy on the estimation of population genetic diversity and the detection of selective loci. *Genetics*, 179(1):539, 2008.
- Carey, VJ, Morgan, M, Falcon, S, et al. Ggtools: analysis of genetics of gene expression in bioconductor. *Bioinformatics*, 23(4):522–523, 2007.
- Cavalli-Sforza, L. Population structure and human evolution. *Proceedings of the Royal Society of London Series B*, 164:362–379, 1966.
- Charlesworth, B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205, 2009.
- Charlwood, J, Kihonda, J, Sama, S, et al. The rise and fall of *Anopheles arabiensis*(Diptera: Culicidae) in a Tanzanian village. *Bulletin of entomological research*, 85(01):37–44, 1995. ISSN 1475-2670.

- Chenet, S, Branch, O, Escalante, A, et al. Genetic diversity of vaccine candidate antigens in *Plasmodium falciparum* isolates from the Amazon basin of Peru. *Malaria journal*, 7(1):93, 2008. ISSN 1475-2875.
- Chiurugwi, T, Beaumont, M, Wilkinson, M, et al. Adaptive divergence and speciation among sexual and pseudoviviparous populations of *Festuca*. *Heredity*, 2010. ISSN 0018-067X.
- Clark, A, Hubisz, M, Bustamante, C, et al. Ascertainment bias in studies of human genome-wide polymorphism. *Genome research*, 15(11):1496, 2005.
- Cock, PJA, Antao, T, Chang, JT, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 2009.
- Cornuet, JM and Luikart, G. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, 144(4):2001–2014, 1996.
- Cortese, JF, Caraballo, A, Contreras, CE, et al. Origin and dissemination of *Plasmodium falciparum* drug-resistance mutations in South America. *J Infect Dis*, 186(7):999–1006, 2002.
- Cosart, T, Beja-Pereira, A, Chen, S, et al. Exome-wide dna capture and sequencing in domestic and wild species. *BMC Genomics*, In press.
- Crofton, J, Chaulet, P, Maher, D, et al. *Guidelines for the management of drug-resistant tuberculosis*. World Health Organization, 1997.
- Cross, A and Singer, B. Modelling the development of resistance of *Plasmodium falciparum* to anti-malarial drugs. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 85(3):349–355, 1991.
- Crow, JF and Kimura, M. *Introduction to Population Genetics Theory*. Harper & Row Publishers, 1970.
- Curtis, CF and Otoo, LN. A simple model of the build-up of resistance to mixtures of anti-malarial drugs. *Trans R Soc Trop Med Hyg*, 80(6):889–892, 1986.
- Curwen, V, Eyra, E, Andrews, TD, et al. The ensembl automatic gene annotation system. *Genome Res*, 14(5):942–950, 2004.
- de Meefis, T and Balloux, F. Clonal reproduction and linkage disequilibrium in diploids: a simulation study. *Infection, Genetics and Evolution*, 4(4):345–351, 2004. ISSN 1567-1348.

- Dondorp, AM, Nosten, F, Yi, P, et al. Artemisinin resistance in *Plasmodium falciparum* malaria. *N Engl J Med*, 361(5):455–467, 2009.
- Doolan, D, Dobano, C, and Baird, J. Acquired immunity to malaria. *Clinical Microbiology Reviews*, 22(1):13, 2009. ISSN 0893-8512.
- Dye, C and Williams, BG. Multigenic drug resistance among inbred malaria parasites. *Proc Biol Sci*, 264(1378):61–67, 1997.
- Dyer, N, Furtado, A, Cano, J, et al. Evidence for a discrete evolutionary lineage within Equatorial Guinea suggests that the tsetse fly *Glossina palpalis palpalis* exists as a species complex. *Molecular Ecology*, 18(15):3268–3282, 2009. ISSN 1365-294X.
- Edwards, A. *Likelihood: an account of the statistical concept of likelihood and its application to scientific inference*, volume 235. Cambridge University Press, 1972.
- Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, 5(6):435–445, 2004. ISSN 1471-0056.
- England, P, Cornuet, JM, Berthier, P, et al. Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conservation Genetics*, 7(2):303–308, 2006.
- England, P, Luikart, G, and Waples, R. Early detection of population fragmentation using linkage disequilibrium estimation of effective population size. *Conservation Genetics*, 11:2425–2430, 2010. ISSN 1566-0621. 10.1007/s10592-010-0112-x.
- Enright, M, Robinson, D, Randle, G, et al. The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). *Proceedings of the National Academy of Sciences of the United States of America*, 99(11):7687, 2002.
- Escalante, A, Grebert, H, Chaiyaroj, S, et al. Polymorphism in the gene encoding the apical membrane antigen-1 (AMA-1) of *Plasmodium falciparum*. X. Asembo Bay Cohort Project. *Molecular and biochemical parasitology*, 113(2):279–287, 2001.
- Excoffier, L and Heckel, G. Computer programs for population genetics data analysis: a survival guide. *Nat Rev Genet*, 7(10):745–758, 2006.
- Excoffier, L, Hofer, T, and Foll, M. Detecting loci under selection in a hierarchically structured population. *Heredity*, 103(4):285–298, 2009. ISSN 0018-067X.
- Excoffier, L, Laval, G, and Schneider, S. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, 1:47–50, 2005.

- Fiore, A, Fry, A, Shay, D, et al. *Antiviral Agents for the Treatment and Chemoprophylaxis of Influenza*. Centers for Disease Control and Prevention, 2011.
- Foll, M and Gaggiotti, O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, 180(2):977, 2008.
- Frankham, R. Effective population size/adult population size ratios in wildlife: a review. *Genetical Research*, 66(02):95–107, 1995. ISSN 1469-5073.
- Frankham, R. Genetics and extinction. *Biological Conservation*, 126(2):131 – 140, 2005. ISSN 0006-3207.
- Gatei, W, Kariuki, S, Hawley, W, et al. Effects of transmission reduction by insecticide-treated bed nets (ITNs) on parasite genetics population structure: I. The genetic diversity of *Plasmodium falciparum* parasites by microsatellite markers in western Kenya. *Malaria Journal*, 9(1):353, 2010.
- Ginger, R, Askew, S, Ogborne, R, et al. SLC24A5 encodes a trans-Golgi network protein with potassium-dependent sodium-calcium exchange activity that regulates human epidermal melanogenesis. *Journal of Biological Chemistry*, 283(9):5486, 2008.
- Goncalves, B and Paul, R. Sub-clearance treatment to slow malaria drug resistance? *Trends Parasitol*, 27:50–51, 2011. ISSN 1471-5007.
- Gregson, A and Plowe, C. Mechanisms of resistance of malaria parasites to antifolates. *Pharmacological reviews*, 57:117–145, 2005. 10.1124/pr.57.1.4.
- Guerin, PJ, Olliaro, P, Nosten, F, et al. Malaria: current status of control, diagnosis, treatment, and a proposed agenda for research and development. *Lancet Infect Dis*, 2(9):564–573, 2002.
- Hastings, I. Why we should effectively treat malaria. *Trends Parasitol*, 27:51–52, 2011.
- Hastings, I and Donnelly, M. The impact of antimalarial drug resistance mutations on parasite fitness, and its implications for the evolution of resistance. *Drug resistance updates*, 8(1-2):43–50, 2005. ISSN 1368-7646.
- Hastings, I, Nsanjabana, C, and Smith, T. A comparison of methods to detect and quantify the markers of antimalarial drug resistance. *The American journal of tropical medicine and hygiene*, 83(3):489, 2010.
- Hastings, I and Watkins, W. Tolerance is the key to understanding antimalarial drug resistance. *Trends in parasitology*, 22(2):71–77, 2006. ISSN 1471-4922.

- Hastings, I, Watkins, W, and White, N. The evolution of drug-resistant malaria: the role of drug elimination half-life. *Phil Trans B*, 357(1420):505, 2002. ISSN 0962-8436.
- Hastings, IM. A model for the origins and spread of drug-resistant malaria. *Parasitology*, 115 (Pt 2):133–141, 1997.
- Hastings, IM. Complex dynamics and stability of resistance to antimalarial drugs. *Parasitology*, 132(Pt 5):615–624, 2006.
- Hastings, IM and D'Alessandro, U. Modelling a predictable disaster: the rise and spread of drug-resistant malaria. *Parasitol Today*, 16(8):340–347, 2000.
- Hastings, IM and Smith, TA. Malhaplofreq: a computer programme for estimating malaria haplotype frequencies from blood samples. *Malar J*, 7:130, 2008.
- Hedrick, PW. Gametic disequilibrium measures: proceed with caution. *Genetics*, 117(2):331–341, 1987.
- Hermisson, J. Who believes in whole-genome scans for selection? *Heredity*, 103(4):283–284, 2009.
- Hill, WG. Estimation of effective population size from data on linkage disequilibrium. *Genetics Research*, 38(03):209–216, 1981.
- Hodges, E, Xuan, Z, Balija, V, et al. Genome-wide in situ exon capture for selective resequencing. *Nature genetics*, 39(12):1522–1527, 2007.
- Hohenlohe, P, Bassham, S, Etter, P, et al. Population genomics of parallel adaptation in threespine stickleback using sequenced rad tags. *PLoS Genetics*, 6(2):e1000862, 2010.
- Howie, B, Donnelly, P, and Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6):e1000529, 2009.
- Hsu, F, Kent, WJ, Clawson, H, et al. The UCSC Known Genes. *Bioinformatics*, 22(9):1036–1046, 2006.
- Hughes, AL and Verra, F. Very large long-term effective population size in the virulent human malaria parasite *Plasmodium falciparum*. *Proc Biol Sci*, 268(1478):1855–1860, 2001.
- Huijben, S, Nelson, W, Wargo, A, et al. Chemotherapy, within-host ecology and the fitness of drug-resistant malaria parasites. *Evolution*, 64:2952–2968, 2010. ISSN 1558-5646.
- Hunter, JD. Matplotlib: A 2d graphics environment. *Computing in Science and Engg*, 9(3):90–95, 2007. ISSN 1521-9615.

- International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, 2007.
- Iwagami, M, Rivera, PT, Villacorte, EA, et al. Genetic diversity and population structure of plasmodium falciparum in the philippines. *Malar J*, 8:96, 2009.
- John, C, Kutamba, E, Mugarura, K, et al. Adjunctive therapy for cerebral malaria and other severe forms of Plasmodium falciparum malaria. *Exp Rev Anti Infect Ther*, 8(9):997–1008, 2010. ISSN 1478-7210.
- Jorde, PE and Ryman, N. Temporal allele frequency change and estimation of effective size in populations with overlapping generations. *Genetics*, 139(2):1077–1090, 1995.
- Jorde, PE and Ryman, N. Demographic genetics of brown trout (*Salmo trutta*) and estimation of effective population size from temporal change of allele frequencies. *Genetics*, 143(3):1369–1381, 1996.
- Jorde, PE and Ryman, N. Unbiased estimator for genetic drift and effective population size. *Genetics*, 177(2):927–935, 2007.
- Kalinowski, S and Waples, R. Relationship of effective to census size in fluctuating populations. *Conservation Biology*, 16:129–136, 2002.
- Kelley, JL, Madeoy, J, Calhoun, JC, et al. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res*, 16(8):980–989, 2006.
- Koella, JC and Antia, R. Epidemiological models for the spread of anti-malarial resistance. *Malar J*, 2:3, 2003.
- Krimbas, CB and Tsakas, S. The Genetics of *Dacus oleae*. V. Changes of Esterase Polymorphism in a Natural Population Following Insecticide Control-Selection or Drift? *Evolution*, 25(3):454–460, 1971.
- Kublin, JG, Cortese, JF, Njunju, EM, et al. Reemergence of chloroquine-sensitive *Plasmodium falciparum* malaria after cessation of chloroquine use in Malawi. *J Infect Dis*, 187(12):1870–1875, 2003.
- Kumar, S and Dudley, J. Bioinformatics software for biologists in the genomics era. *Bioinformatics*, 23(14):1713–1717, 2007.
- Langhorne, J, Ndungu, FM, Sponaas, AM, et al. Immunity to malaria: more questions than answers. *Nat Immunol*, 9(7):725–732, 2008.
- Laufer, MK, Thesing, PC, Eddington, ND, et al. Return of chloroquine antimalarial efficacy in Malawi. *N Engl J Med*, 355(19):1959–1966, 2006.

- Leberg, P. Genetic approaches for estimating the effective size of populations. *Journal of Wildlife Management*, 69(4):1385–1399, 2005. ISSN 0022-541X.
- Lehmann, T, Hawley, W, Grebert, H, et al. The effective population size of *Anopheles gambiae* in Kenya: implications for population structure. *Molecular Biology and Evolution*, 15(3):264, 1998. ISSN 0737-4038.
- Lewontin, RC and Krakauer, J. Letters to the editors: Testing the heterogeneity of F values. *Genetics*, 80(2):397–398, 1975.
- Looareesuwan, S, Viravan, C, Webster, HK, et al. Clinical studies of atovaquone, alone or in combination with other antimalarial drugs, for treatment of acute uncomplicated malaria in Thailand. *Am J Trop Med Hyg*, 54(1):62–66, 1996.
- Luikart, G, Cornuet, JM, and Allendorf, FW. Temporal changes in allele frequencies provide estimates of population bottleneck size. *Conservation Biology*, 13(3):523–530, 1999. ISSN 08888892.
- Luikart, G, England, PR, Tallmon, D, et al. The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet*, 4(12):981–994, 2003.
- Luikart, G, Ryman, N, Tallmon, D, et al. Estimating census and effective population sizes: Increasing usefulness of genetic methods. *Conservation Genetics*, in press.
- Luikart, G, Sherwin, WB, Steele, BM, et al. Usefulness of molecular markers for detecting population bottlenecks via monitoring genetic change. *Molecular Ecology*, 7(8):963–974, 1998.
- Matukumalli, L, Lawley, C, Schnabel, R, et al. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One*, 4(4):e5350, 2009.
- Maude, R, Pontavornpinyo, W, Saralamba, S, et al. The last man standing is the most resistant: eliminating artemisinin-resistant malaria in Cambodia. *Malaria J*, 8(1):31, 2009. ISSN 1475-2875.
- Mishra, S and Newton, C. Diagnosis and management of the neurological complications of *falciparum* malaria. *Nat Rev Neurol*, 5(4):189–198, 2009. ISSN 1759-4758.
- Morin, P, Luikart, G, and Wayne, R. SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, 19(4):208–216, 2004.
- Mzilahowa, T, McCall, PJ, and Hastings, IM. "sexual" population structure and genetics of the malaria agent *p. falciparum*. *PLoS One*, 2(7):e613, 2007.
- Narum, S and Hess, JE. Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources*, 11:184–194, 2011. ISSN 1755-0998.

- Nei, M and Tajima, F. Genetic drift and estimation of effective population size. *Genetics*, 98(3):625–640, 1981.
- Nielsen, R. Molecular signatures of natural selection. *Annual Reviews in Genetics*, 39:197–218, 2005.
- Nomura, T. Estimation of effective number of breeders from molecular coancestry of single cohort sample. *Evolutionary Applications*, 1(3):462–474, 2008.
- Nunney, L and Elam, DR. Estimating the effective population size of conserved populations. *Conservation Biology*, 8:175–184, 1994.
- Oliphant, TE. *Guide to NumPy*. Provo, UT, 2006.
- Olliaro, P. Mode of action and mechanisms of resistance for antimalarial drugs. *Pharmacol Ther*, 89(2):207–219, 2001.
- Olliaro, P. Drug resistance hampers our capacity to roll back malaria. *Clin Infect Dis*, 41 Suppl 4:S247–S257, 2005.
- O’Ryan, C, Harley, E, Bruford, M, et al. Microsatellite analysis of genetic diversity in fragmented South African buffalo populations. *Animal Conservation*, 1(02):85–94, 1998. ISSN 1469-1795.
- Osman, ME, Mockenhaupt, FP, Bienzle, U, et al. Field-based evidence for linkage of mutations associated with chloroquine (PfCRT/PfMDR1) and sulfadoxine-pyrimethamine (PfDHFR/PfDHPS) resistance and for the fitness cost of multiple mutations in *P. falciparum*. *Infect Genet Evol*, 7(1):52–59, 2007.
- Otto, S. The evolutionary enigma of sex. *The American Naturalist*, 174:1–14, 2009. ISSN 0003-0147.
- Otto, SP and Lenormand, T. Resolving the paradox of sex and recombination. *Nature Review Genetics*, 3(4):252–261, 2002.
- Ovenden, J, Peel, D, Street, R, et al. The genetic effective and adult census size of an Australian population of tiger prawns (*Penaeus esculentus*). *Molecular Ecology*, 16(1):127–138, 2007.
- Palstra, FP and Ruzzante, DE. Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? *Molecular Ecology*, 2008.
- Pearce, RJ, Drakeley, C, Chandramohan, D, et al. Molecular determination of point mutation haplotypes in the dihydrofolate reductase and dihydropteroate synthase of *Plasmodium falciparum* in three districts of northern Tanzania. *Antimicrob Agents Chemother*, 47(4):1347–1354, 2003.

- Peng, B and Kimmel, M. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18):3686–3687, 2005.
- Pérez-Figueroa, A, García-Pereira, M, Saura, M, et al. Comparing three different methods to detect selective loci using dominant markers. *Journal of Evolutionary Biology*, 23(10):2267–2276, 2010. ISSN 1420-9101.
- Pinto, J, Donnelly, M, Sousa, C, et al. Genetic structure of *Anopheles gambiae* (*Diptera: Culicidae*) in Sao Tome and Principe (West Africa): implications for malaria control. *Molecular Ecology*, 11(10):2183–2187, 2002.
- Pinto, J, Donnelly, M, Sousa, C, et al. An island within an island: genetic differentiation of *Anopheles gambiae* in Sao Tome, West Africa, and its relevance to malaria vector control. *Heredity*, 91(4):407–414, 2003. ISSN 0018-067X.
- Plowe, C, Cortese, J, Djimde, A, et al. Mutations in *Plasmodium falciparum* dihydrofolate reductase and dihydropteroate synthase and epidemiologic patterns of pyrimethamine-sulfadoxine use and resistance. *The Journal of infectious diseases*, 176:1590–1596, 1997.
- Plowe, CV, Kublin, JG, Dzinjalimala, FK, et al. Sustained clinical efficacy of sulfadoxine-pyrimethamine for uncomplicated *falciparum* malaria in Malawi after 10 years as first line treatment: five year prospective study. *BMJ*, 328(7439):545, 2004.
- Pollak, E. A new method for estimating the effective population size from allele frequency changes. *Genetics*, 104(3):531–548, 1983.
- Pongtavornpinyo, W, Hastings, IM, Dondorp, A, et al. Probability of emergence of antimalarial resistance in different stages of the parasite life cycle. *Evolutionary Applications*, 2:52–61(10), 2009.
- Price, RN, Uhlemann, AC, Brockman, A, et al. Mefloquine resistance in *Plasmodium falciparum* and increased PFMDR1 gene copy number. *Lancet*, 364(9432):438–447, 2004.
- R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.
- Raymond, M and Rousset, F. GENEPOP: population genetics software for exact tests and ecumenicism. *Journal of Heredity*, 86:248–249, 1995.
- Rhead, B, Karolchik, D, Kuhn, RM, et al. The UCSC Genome Browser database: update 2010. *Nucleic Acids Res*, 38(Database issue):D613–D619, 2010.

- Richards, C and Leberg, PL. Temporal changes in allele frequencies and a population's history of severe bottlenecks. *Conservation Biology*, 10(3):832–839, 1996. ISSN 08888892.
- Rios, D, McLaren, WM, Chen, Y, et al. A database and API for variation, dense genotyping and resequencing data. *BMC Bioinformatics*, 11:238, 2010.
- Robertson, A. The interpretation of genotypic ratios in domestic animal populations. *Animal Production*, 7(03):319–324, 1965.
- Rogers, W, Sem, R, Tero, T, et al. Failure of artesunate-mefloquine combination therapy for uncomplicated *Plasmodium falciparum* malaria in southern Cambodia. *Malaria Journal*, 8(1):10, 2009. ISSN 1475-2875.
- Rogerson, S, Wijesinghe, R, and Meshnick, S. Host immunity as a determinant of treatment outcome in *Plasmodium falciparum* malaria. *The Lancet Infectious Diseases*, 10(1):51–59, 2010. ISSN 1473-3099.
- Roper, C, Pearce, R, Nair, S, et al. Intercontinental spread of pyrimethamine-resistant malaria. *Science*, 305(5687):1124, 2004.
- Rousset, F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, 8(1):103–106, 2008.
- Sa, J and Twu, O. Protecting the malaria drug arsenal: halting the rise and spread of amodiaquine resistance by monitoring the PfCRT SVMNT type. *Malaria Journal*, 9(1):374, 2010. ISSN 1475-2875.
- Schug, MD, Mackay, TF, and Aquadro, CF. Low mutation rates of microsatellite loci in *Drosophila melanogaster*. *Nature Genetics*, 15(1):99–102, 1997.
- Schwartz, M, Tallmon, D, and Luikart, G. Using genetics to estimate the size of wild populations: many methods, much potential, uncertain utility. *Animal Conservation*, 2(04):321–323, 1999.
- Schwenke, A, Brandts, C, Philipps, J, et al. Declining chloroquine resistance of *Plasmodium falciparum* in Lambaréné, Gabon from 1992 to 1998. *Wien Klin Wochenschr*, 113(1-2):63–64, 2001.
- Sibley, CH, Hyde, JE, Sims, PF, et al. Pyrimethamine-sulfadoxine resistance in *Plasmodium falciparum*: what next? *Trends Parasitol*, 17(12):582–588, 2001.
- Simard, F, Lehmann, T, Lemasson, J, et al. Persistence of *Anopheles arabiensis* during the severe dry season conditions in Senegal: an indirect approach using microsatellite loci. *Insect Molecular Biology*, 9(5):467–479, 2000. ISSN 1365-2583.

- Simmons, L. *Sperm competition and its evolutionary consequences in the insects*. Princeton Univ Pr, 2001. ISBN 0691059888.
- Slominski, A, Tobin, D, Shibahara, S, et al. Melanin pigmentation in mammalian skin and its hormonal regulation. *Physiological reviews*, 84(4):1155, 2004.
- Smedley, D, Haider, S, Ballester, B, et al. BioMart—biological queries made easy. *BMC Genomics*, 10:22, 2009.
- Smith, T, Killeen, G, Maire, N, et al. Mathematical modeling of the impact of malaria vaccines on the clinical epidemiology and natural history of *Plasmodium falciparum* malaria: Overview. *The American Journal of Tropical Medicine and Hygiene*, 75(2 suppl):1, 2006.
- Smith, T, Maire, N, Ross, A, et al. Towards a comprehensive simulation model of malaria epidemiology and control. *Parasitology*, 135(13):1507–1516, 2008. ISSN 1469-8161.
- Snow, R, Guerra, C, Noor, A, et al. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature*, 434:214–217, 2005. 10.1038/nature03342.
- Sokal, RR and Rohlf, FJ. *Biometry*. W. H. Freeman and Co.: New York, 1995. ISBN 0-7167-2411-1.
- Solano, P, Ravel, S, Bouyer, J, et al. The population structure of *Glossina palpalis gambiensis* from island and continental locations in coastal Guinea. *PLoS Negl Trop Dis*, 3(3):e392, 2009.
- SQLite Development team. The SQLite database engine. 2010.
- Stajich, JE, Block, D, Boulez, K, et al. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–1618, 2002.
- Stein, LD, Mungall, C, Shu, S, et al. The generic genome browser: a building block for a model organism system database. *Genome Res*, 12(10):1599–1610, 2002.
- Storz, JF. Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol Ecol*, 14(3):671–688, 2005.
- Susomboon, P, Iwagami, M, Tangpukdee, N, et al. Differences in genetic population structures of *Plasmodium falciparum* isolates from patients along Thai-Myanmar border with severe or uncomplicated malaria. *Malar J*, 7:212, 2008.
- Talisuna, AO, Langi, P, Bakyaite, N, et al. Intensity of malaria transmission, antimalarial-drug use and resistance in Uganda: what is the relationship between these three factors? *Trans R Soc Trop Med Hyg*, 96(3):310–317, 2002.

- Tallmon, D, Gregovich, D, Waples, R, et al. When are genetic methods useful for estimating contemporary abundance and detecting population trends? *Molecular Ecology Resources*, 10(4):684–692, 2010. ISSN 1755-0998.
- Tallmon, DA, Koyuk, A, Luikart, G, et al. onesamp: a program to estimate effective population size using approximate bayesian computation. *Molecular Ecology Notes*, 8:299–301(3), 2008.
- Targett, G and Greenwood, B. Malaria vaccines and their potential role in the elimination of malaria. *Malaria J*, 7(Suppl 1):S10, 2008. ISSN 1475-2875.
- Taylor, C, Toure, Y, Coluzzi, M, et al. Effective population size and persistence of *Anopheles arabiensis* during the dry season in West Africa. *Medical and veterinary entomology*, 7(4):351–357, 1993. ISSN 1365-2915.
- Tediosi, F, Maire, N, Smith, T, et al. An approach to model the costs and effects of case management of *Plasmodium falciparum* malaria in sub-Saharan Africa. *The American Journal of Tropical Medicine and Hygiene*, 75(2 suppl):90, 2006.
- The 1000 genomes project team. The 1000 genomes project. 2010.
- Thorisson, GA, Smith, AV, Krishnan, L, et al. The International HapMap Project Web site. *Genome Res*, 15(11):1592–1593, 2005.
- Tishkoff, SA, Reed, FA, Ranciaro, A, et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet*, 39(1):31–40, 2007.
- Tripet, F, Toure, Y, Taylor, C, et al. DNA analysis of transferred sperm reveals significant levels of gene flow between molecular forms of *Anopheles gambiae*. *Molecular Ecology*, 10(7):1725–1732, 2001. ISSN 0962-1083.
- Turner, T, Salter, L, and Gold, J. Temporal-method estimates of Ne from highly polymorphic loci. *Conservation Genetics*, 2(4):297–308, 2001.
- Tuteja, R. Malaria – an overview. *FEBS Journal*, 274:4670–4670, 2007. 10.1111/j.1742-4658.2007.05997.x.
- Ursing, J, Schmidt, BA, Lebbad, M, et al. Chloroquine resistant *P. falciparum* prevalence is low and unchanged between 1990 and 2005 in Guinea-Bissau: an effect of high chloroquine dosage? *Infect Genet Evol*, 7(5):555–561, 2007.
- Vaidya, A and Mather, M. Atovaquone resistance in malaria parasites. *Drug Resist Updat*, 3(5):283–287, 2000.
- Vinayak, S, Alam, M, Mixson-Hayden, T, et al. Origin and evolution of sulfadoxine resistant *Plasmodium falciparum*. *PLoS Pathogens*, 6(3):e1000830, 2010.

- Vitalis, R and Couvet, D. Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics*, 157(2):911–925, 2001.
- Vitalis, R, Dawson, K, Boursot, P, et al. DetSel 1.0: a computer program to detect markers responding to selection. *Journal of Heredity*, 94(5):429, 2003.
- Walton, C, Handley, J, et al. Population structure and population history of *Anopheles dirus* mosquitoes in Southeast Asia. *Molecular Biology and Evolution*, 17(6):962, 2000. ISSN 0737-4038.
- Wang, J. A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetics Research*, 78(3):243–257, 2001.
- Wang, J and Whitlock, MC. Estimating effective population size and migration rates from genetic samples over space and time. *Genetics*, 163(1):429–446, 2003.
- Waples, R and Do, C. Linkage disequilibrium estimates of contemporary N_e using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications*, 3(3):244–262, 2010. ISSN 1752-4571.
- Waples, R and Gaggiotti, O. What is a population?: An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular ecology*, 15:1419–1439, 2006. 10.1111/j.1365-294X.2006.02890.x.
- Waples, RS. A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics*, 121(2):379–391, 1989.
- Waples, RS. A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci*. *Conservation Genetics*, 7(2):167–184, 2006.
- Waples, RS and Do, C. ldne: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources*, 8(4):753–756, 2008.
- Waples, RS and Yokota, M. Temporal estimates of effective population size in species with overlapping generations. *Genetics*, 175(1):219–233, 2007.
- Wargo, AR, Huijben, S, de Roode, JC, et al. Competitive release and facilitation of drug-resistant parasites after therapeutic chemotherapy in a rodent malaria model. *Proc Natl Acad Sci U S A*, 104(50):19914–19919, 2007.
- Weetman, D, Wilding, C, Steen, K, et al. Association Mapping of Insecticide Resistance in Wild *Anopheles gambiae* Populations: Major Variants Identified in a Low-Linkage Disequilibrium Genome. *PloS one*, 5(10):e13140, 2010. ISSN 1932-6203.
- Wehrle-Haller, B. The Role of Kit-Ligand in Melanocyte Development and Epidermal Homeostasis. *Pigment cell research*, 16(3):287–296, 2003.

- Weir, BS and Cockerham, CC. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6):1358–1370, 1984.
- Wellems, TE and Plowe, CV. Chloroquine-resistant malaria. *J Infect Dis*, 184(6):770–776, 2001.
- White, NJ and Pongtavornpinyo, W. The de novo selection of drug-resistant malaria parasites. *Proc Biol Sci*, 270(1514):545–554, 2003.
- Wondji, C, Simard, F, Lehmann, T, et al. Impact of insecticide-treated bed nets implementation on the genetic structure of *Anopheles arabiensis* in an area of irrigated rice fields in the Sahelian region of Cameroon. *Molecular Ecology*, 14(12):3683–3693, 2005. ISSN 1365-294X.
- Wootton, J, Feng, X, Ferdig, M, et al. Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature*, 418(6895):320–323, 2002.
- World Health Organization. *Guidelines for the treatment of malaria*. World Health Organization, 2006.
- World Health Organization. *World malaria report 2008*. WHO, 2008. ISBN 9241563699.
- Wright, S. Evolution in mendelian populations. *Genetics*, 16(2):97–159, 1931.
- Yang, H, Yang, Y, Yang, P, et al. Monitoring *Plasmodium falciparum* chloroquine resistance in Yunnan Province, China, 1981-2006. *Acta Trop*, 108(1):44–49, 2008.
- Yeung, S, Pongtavornpinyo, W, Hastings, IM, et al. Antimalarial drug resistance, artemisinin-based combination therapy, and the contribution of modeling to elucidating policy choices. *Am J Trop Med Hyg*, 71(2 Suppl):179–186, 2004.
- Zhivotovsky, L. Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology*, 8(6):907–913, 1999. ISSN 1365-294X.
- Zignol, M, Hosseini, M, Wright, A, et al. Global incidence of multidrug-resistant tuberculosis. *Journal of Infectious Diseases*, 194(4):479, 2006. ISSN 0022-1899.

Appendices

A

ogaraK: Supplementary material

A.1 Simulator overview

OgaraK simulates a *P. falciparum* population which suffers heterogeneous selection pressure over time and space. Here we present the model with a special focus on the biological underpinnings and some of the more uncommon properties (i.e. either not captured by existing simulators or at least not commonly simulated) are emphasised.

OgaraK tracks the spread of drug resistance using a time-scale of parasite generations (a generation is a parasite reproduction cycle from host to host), where resistance to each drug is encoded by one or more loci. The main output of the simulator is the frequency of resistance genotypes over time.

Biological observations

While we do not intend to present an introduction to *P. falciparum* biology (see e.g., Tuteja (2007)), some observations, especially those that differ from standard population genetics, are important in order to understand the main design options of ogaraK:

Plasmodium malaria parasites are haploid and reproduce asexually within humans. Infections can include up to 10^{12} parasites. They undergo a brief diploid phase with sexual recombination in the mosquito.

Humans often contain several genetically distinct *P. falciparum* clones acquired from different mosquito bites; the number of clones in a human is called the multiplicity of infection (MOI).

For most drug treatments, resistance only emerged in a small number of foci (Wellems and Plowe, 2001; Roper et al., 2004), the notable exception being Atovaquone (Vaidya and Mather, 2000). While understanding the appearance of *de novo* mutations is an important topic, (see e.g., White and Pongtavornpinyo (2003); Pongtavornpinyo et al. (2009)), in the vast majority of human populations, resistance will not occur via mutation, but via immigration of resistance genes or is already present for many drugs.

While inside the mosquito, mating options are severely restricted to the gametes coming from the blood meal, and mating between genetic equal gametes is not uncommon.

mon. Mating between gametes from the same clone (selfing) results in sex between identical haploid genotypes resulting in clonal reproduction. Mating between different clones (out-crossing) results in genetic recombination and re-assortment of *P. falciparum* genes. If a parasite has a mutation and is in competition with wild-type variant on an untreated individual, a fitness penalty is normally incurred by the resistant parasites (Babiker et al., 2009). While the fitness penalty probably happens while in direct competition with other infections in the blood stage, the hypothesis of occurring elsewhere (e.g. liver stage, or while in transmission) cannot be realistically excluded.

Genetically distinct gametes ingested by a mosquito are therefore related to the MOI: the number of distinct genotypes that can be ingested cannot be bigger than the MOI. MOI is a proxy for the intensity of transmission: in areas of high transmission the MOI is bigger due to repeated sequential infection. Typical MOI values are below 7 with higher values occurring in areas of intense transmission.

Humans might develop partial immunity to malaria over time (Langhorne et al., 2008), this allows for asymptomatic carriers which might be untreated (i.e., without selection pressure imposed by a drug). Existing evidence (Osman et al., 2007) suggests that most, if not all, loci involved in drug resistance are physically unlinked.

From these biological insights, ogaraK is designed along the following lines:

- The parasite population is defined as the population co-existing with a human host population. Different hosts (humans) provide different selection environments, e.g., untreated hosts act as potential reservoirs for sensitive parasites and treated hosts will kill sensitive forms while resistant infections may survive treatment. Each type of environment contains parasites facing different selection pressures.
- Each infection has a certain genetic profile. For each drug a certain number of loci are involved in drug resistance. Each locus has two alleles: wild (“sensitive”) and mutant (“resistant”). The ability to resist drug treatment is dependent on the relationships between resistant loci and is detailed below.
- The demographic model presented is equivalent to an island model with an extreme migration rate proportional to the relative size of each environment, making the environment to which a parasite is exposed independent of the environment of the parent(s). We note that, in most standard demographic models – e.g., island or stepping-stone – the migration rate is normally low and there is a large probability of an offspring staying within the population of the parents. That is not the case here, but strictly speaking we implement an island model.
- Mutated parasites are expected to exist at the onset of the simulation, but we also support mutation (*de novo* emergence).

- Individual simulation of parasites is impossible because of large population sizes. The model used to perform the simulation is detailed below.
- While inside human individuals *P. falciparum* parasites are only in haploid form, therefore no effort is made to model diploidy (e.g., dominance effects) inside humans. Sexual forms (gametocytes) do exist in the human, but these are still haploid.
- The level of parasite inbreeding is directly influenced by the MOI parameter, as the mating options inside the mosquito are mostly influenced by the diversity available in a single blood meal coming from a single individual.
- All loci are modelled as physically unlinked.
- The fitness penalty can be modelled as blood-stage (erythrocytic) or non-blood stage (exo-erythrocytic).

A very simple example is presented on figure A.1. Here a single drug is deployed and given to 40% of the population. This means a single selection pressure outside untreated individuals, that does not change over time (i.e., the treatment rate is constant). This simple example is mainly provided as a basis for a build-up for more complex examples provided below.

With the design options presented above, ogaraK is able to approximate the vast majority of the literature using a population genetics approach modelling the spread of drug resistance.

Epistasis

Resistance to drugs often involves more than one locus. Modes of resistance are not fully understood for most drugs, with notable exception of SP (Gregson and Plowe, 2005). In order to understand the importance of gene interactions when more than one locus is involved, ogaraK allows for specification of the epistasis mode among loci. Three modes are available: Full epistasis (requiring all loci to be mutant in order for a parasite to be able to resist), duplicate gene function (DGF - only one locus is necessary to confer resistance) and asymmetric epistasis (one locus is necessary, while a second is irrelevant). Asymmetric epistasis is only invoked in conjunction with other modes (see below) and obviously is no different from a single-locus model if used on its own. Asymmetry is inspired from the modes of resistance against Chloroquine and SP when a “more important” locus (*pfert* in the case of CQ and *pfdhfr* in the case of SP) is complemented by a “less important” one (*pfmdr* on CQ and *pfdhps* on SP). A summary of supported epistasis modes is presented on table A.1. Interestingly, the most commonly modelled epistasis mode in the literature (full-epistasis) does not seem to be the most widely reported from empirical studies (like e.g., the CQ and SP cases reported above).

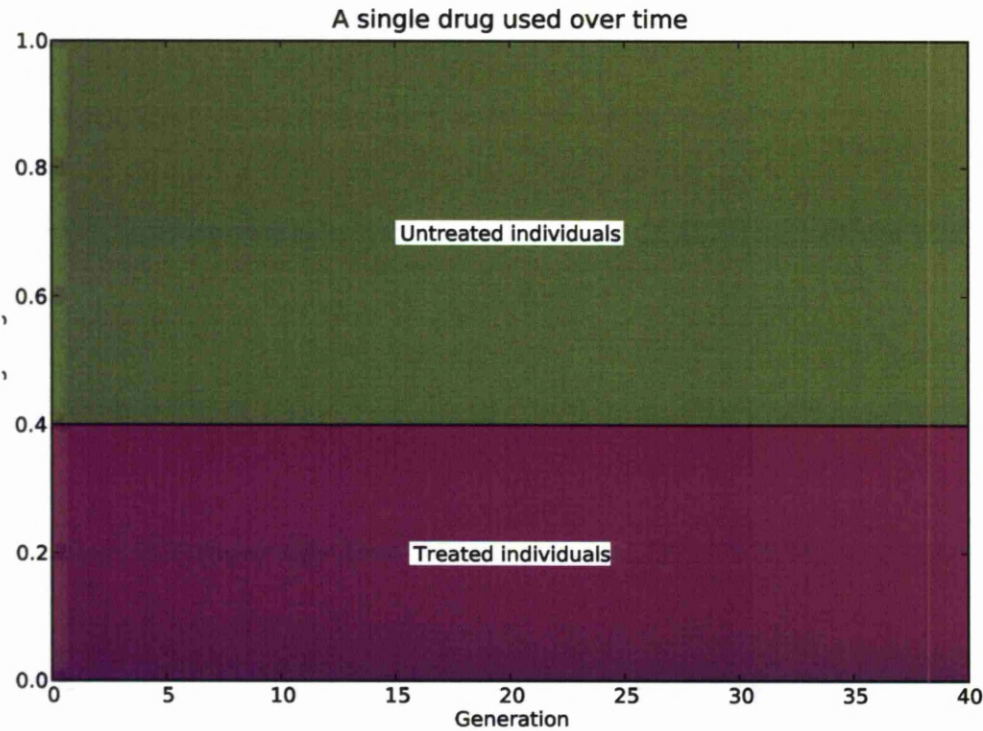


Figure A.1: One drug is used over 40 generations. 40% of individuals are treated, while the remainder remain untreated. Only a single drug is used over time and nothing is said about the genetic profile of resistance.

Resistance profile		Epistasis mode		
Locus 1	Locus 2	Full	Asymmetric	DGF
Sens	Sens	Cure	Cure	Cure
Sens	Res	Cure	Cure	Resistance
Res	Sens	Cure	Resistance	Resistance
Res	Res	Resistance	Resistance	Resistance

Table A.1: The outcome of human drug treatment according to genetic modes of resistance. A single mutation at either loci is enough to confer resistance with duplicate gene function (DGF). In asymmetric epistasis, the first locus is necessary and sufficient to confer resistance. Mutations at both loci are necessary with full epistasis.

Simulations can therefore be done with different epistasis modes in order to understand the effect of different gene interactions. Most importantly different epistasis modes can be used in the same simulation in order to simulate different human environments. The reason we define alternate forms of epistasis is that “resistance” is not an intrinsic trait of the parasites’ genotypes, but depends on the human “environment” in which

treatment occurs. For example: Parasites may require mutations at both loci (i.e. full epistasis) to survive treatment in humans who take a full drug course and have some anti-malaria immunity; parasites may require mutations at only the main locus (i.e. asymmetric epistasis) to survive treatment in humans who are non-immune or who take an incomplete drug course; parasites may require only a single mutation at either locus (duplicate gene function) to survive treatment in humans who are non-immune and do not take a complete drug course. It is in such context that asymmetric epistasis can be useful: While in some human hosts the most important locus is enough to confer resistance, in others both loci are required for parasite survival and parasites “jump” among different kinds of hosts during the simulation. This can have implications in linkage disequilibrium patterns between resistance loci (Antao and Hastings, 2011b). An example scenario is presented on figure A.2.

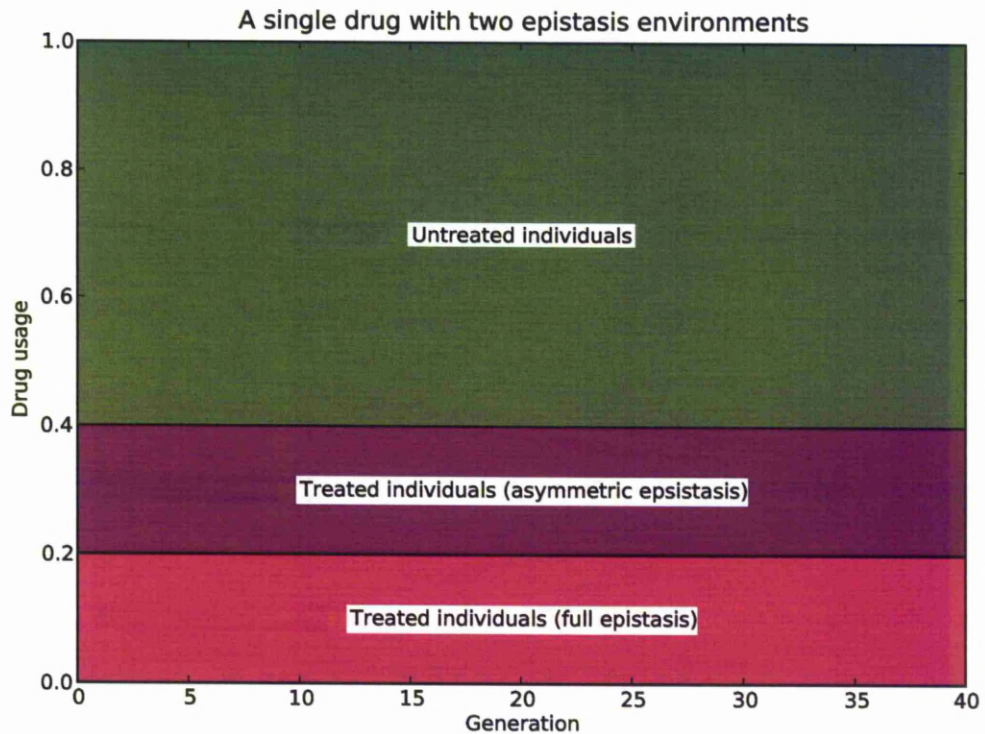


Figure A.2: One drug is used over 40 generations. 40% of individuals are treated, while the remainder remain untreated. Only a single drug is used over time. Half of the treated population takes the full treatment course and thus is able to clear all parasites except the ones having all mutations related to the drug used. The other half of the population, fail to take the full course so are only able to clear the sensitive parasites and the infections having only the “least important” mutation in the asymmetric model.

Mathematical formalisation

In order to simplify the presentation of the mathematical formalisation, we describe a single case scenario where 2 drugs are deployed and 2 loci per drug are used to code for resistance (the drugs will be non-ACT, i.e., they will not share any resistance loci). MOI will be set constant at 2 and the fitness penalty will be modelled at the blood stages. A more general formalisation (though not including epistasis) can be seen in Hastings (2006).

We note that there are 16 possible genotypes, from a totally sensitive one (genotype 0 – no mutations) to the multiple-resistant one (genotype 15 – all mutations). The resistance profile can be easily computed by imagining the binary representation of a clone where the first drug is represented at the right hand side. Clones 1 (0001) and 2 (0010) have one mutation related to drug one, clone 3 (0011) has both. Clone 13 (1101) has both mutations related to drug two and one related to drug one.

The frequency of each resistance profile k transmitted to the next generation will be:

$$F'_k = \frac{\sum_{c_1=0}^{15} \sum_{c_2=0}^{15} \sum_{e=0}^2 \sum_{d=0}^2 f(e)p(c_1, c_2)p(d)t_{e,d}(k, c_1, c_2)}{\bar{W}} \quad (\text{A.1})$$

The structure of the equation can be explained as follows: All possible combinations of clones (as MOI is 2 we consider combinations of two, but any number of clones can be combined using a multinomial distribution) are considered on the summations of c_1 and c_2 . Summation of e allows to investigate the contribution of all environments (untreated individuals plus one or more different epistasis cases). $f(e)$ is the fraction of the host population providing a certain environment (e.g., untreated, treated with duplicate gene function, etc.). $p(d)$ is the probability of receiving a certain drug regimen where $d = 0$ means no drug. For the untreated environment $p(0) = 1$. For other environments $p(0) = 0$ and $p(d > 0)$ is dependent on policy (e.g., $\frac{1}{n_d}$ with MFT). For each environment e there is a different transmission proportion for each profile ($t_{e,d}$, see below).

\bar{W} is a normalisation coefficient equal to the sum of all the numerators (in order to assure that the proportions of each type of transmission $\sum_{k=0}^{2^l-1} F'_k$ sum to 1).

Supporting more than one MOI per simulation is done by simply adding similar parcels in proportion to the frequency of a certain MOI (details can be seen in Hastings (2006)).

The transmission is dependent on the drug applied and epistasis. Two concrete examples will make this more clear:

1. If no drug is present and clones 0 (0000) and 3 (0011) mate (i.e., no mutations and the two mutations involved in resistance to the first drug) then clones 0, 1, 2 and 3 might be transmitted. Note that clones 1 and 2 are possible to emerge via

recombination. Because of fitness penalties clone 0 will have a bigger probability of transmitting than 1, 2 and 3. Clones 1 and 2 will have equal probabilities (same number of mutations). Clone 3 will be the least probable to transmit.

2. If the second drug is present in an environment with full epistasis and clones 12 (1100) and 4 (0100) are present (clone 12 has both mutations for the second drug and clone 4 only one). In this case, the drug eliminates clone 4. Only clone 12 transmits with no recombination possible.

A.2 Drug policies

Three drug policies are supported, causing qualitatively different pressure types:

Rotation A drug is replaced by another every n generations to avoid resistance spreading to high levels. Reintroduction is possible after the whole arsenal of drugs is used once (figure A.3a). Drugs can be rotated not only every n generations, but after resistance to the current drug reaches a certain threshold.

Multiple first line therapies (MFT) All drugs are used simultaneously, but different individuals are treated with a single drug (figure A.3b).

Combination therapy Each individual is treated with all the drugs available, this is akin to single drug usage when the number of loci involved in resistance to the combination cocktail is the sum of all loci involved in each drug.

Drug policies can be combined with different epistasis modes, allowing for quite complex scenarios with both spatial and temporal heterogeneity patterns, as an example, figure A.4 depicts a scenario where two drugs are deployed in sequence and each drug has two epistasis modes.

Artemisinin combination therapies

Artemisinin combination therapies (ACTs) are the current WHO recommended treatment (World Health Organization, 2006) and MFT and rotation of this type of drugs provides a novel model of genetic interaction between drugs. ACTs, as its name implies involve the combination of an Artemisinin derivative with a partner drug. This means that, if resistance surfaces, it will probably involve both the Artemisinin derivative and the partner drug. The Artemisinin derivative is present in all ACTs, so it is necessary to model resistance to ACTs by allowing the genetic mode of Artemisinin resistance to be shared among different treatments. ACT drugs always have two loci per drug, but one of the loci is shared among all drugs. As such, if 3 ACT drugs are modelled, 4 loci are tracked: one for each partner drug, and one representing the Artemisinin derivative shared by all.

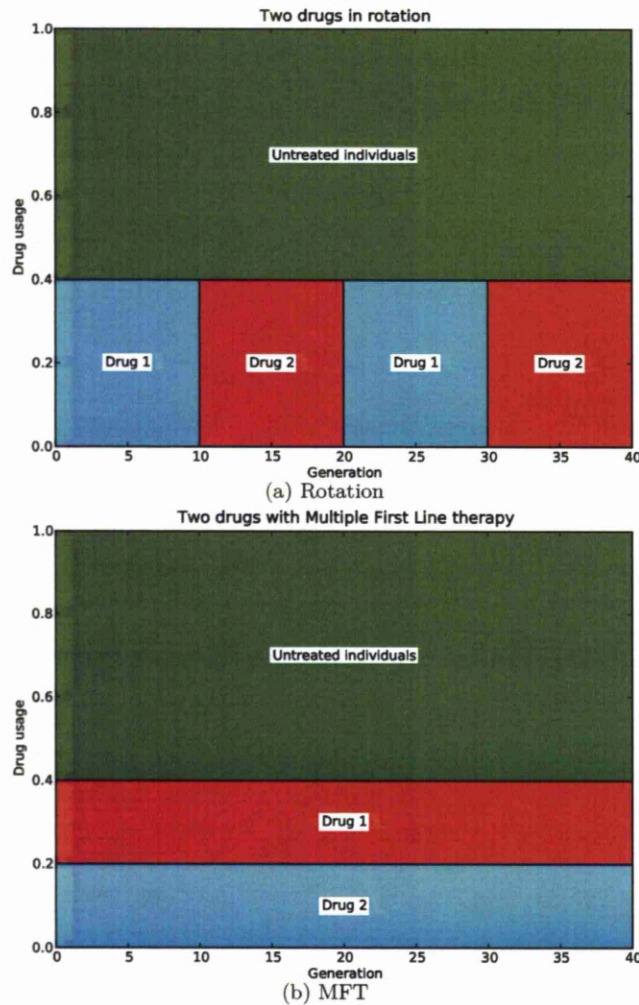


Figure A.3: Different drug deployment policies. Rotation entails swapping drugs every n generations, MFT using all drugs simultaneously, but only a single drug per individual. Different drugs have different resistance mechanisms (i.e. loci).

Mutation

OgaraK main focus is the study of the spread of resistance, but *de novo* emergence is also supported.

OgaraK has no concept of the number of parasites, as it tracks infections. Mutation is therefore supported using two parameters: the probability of a mutation occurring per generation (applicable to each locus independently) and the percentage of sensitive clones which are mutated to resistant. The following important points are noted:

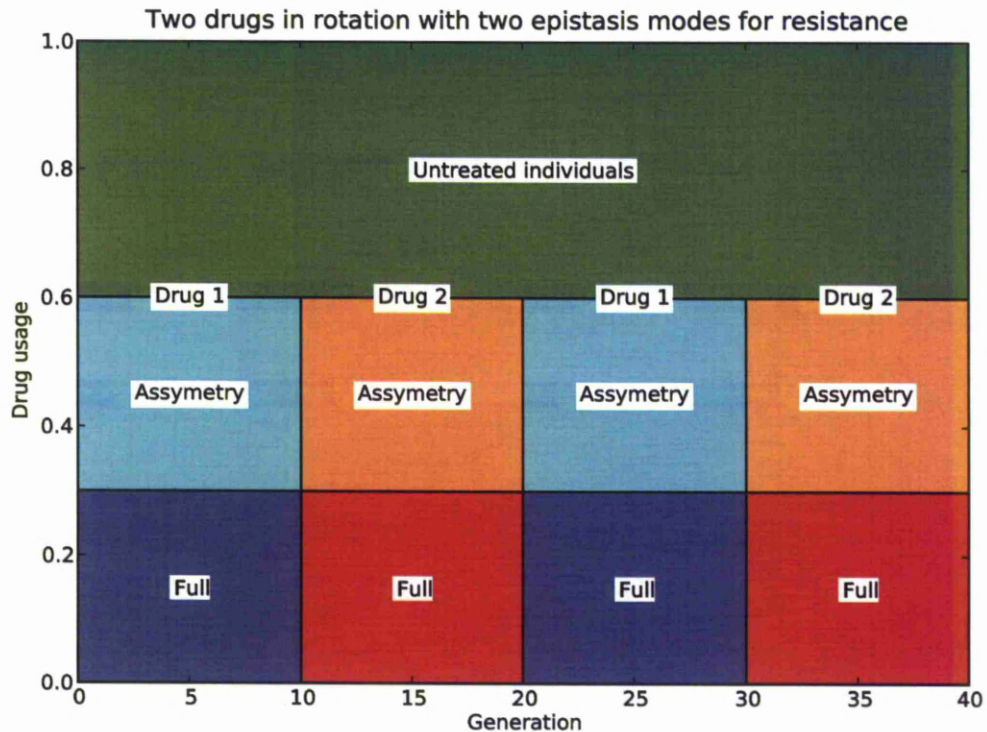


Figure A.4: Multiple epistasis modes can be combined with different policies. In this example, 2 drugs are used with a rotation policy, each drug has two different epistasis modes to account for different immunity profiles.

1. We track the conversion of sensitive clones to resistant, but not the other way around as such event is not relevant to malaria epidemiology. Even when a drug is removed, sensitive clones re-appear from existing sensitive parasites as, in reality there is at least residual frequencies of sensitive infections.
2. The concept of *de novo* mutation occurring rarely (i.e. not in every generation) is firmly grounded in empirical data for the emergence of resistance in most drugs.
3. All mutations are independent and all loci have the same probability of mutation.
4. If a mutation occurs, it can happen, randomly, to any clone which is sensitive at the mutated locus. A percentage of that clone (specified as a parameter) will be converted to a resistant clone.

A fundamental point should be stressed: If mutation is activated, ogaraK which is normally a deterministic simulator, becomes stochastic. While for studies of spread of resistance (no mutation) the results do not vary with parameters, if mutation is activated different runs for the same parameters might have held different results.

Parallels with sex theory

While ogaraK was not designed with sex theory (Otto and Lenormand, 2002) as its first use case, the conceptual parallels between drug rotation and the Red Queen Hypothesis and between MFT and spatial heterogeneity are clear and therefore it can be used to test some (but not all) existing theories for the evolution of sex and recombination. In general, ogaraK can be used to easily simulate very heterogeneous selection patterns.

A.3 Simple user guide

OgaraK is a population genetics simulator, this means that it can be used to generate synthetic datasets but not to do data analysis (though we supply many examples of example scripts in the Python language). The data that is generated is the frequency of genotypes in the parasite population each generation. OgaraK greatly facilitates the data analysis procedure by generating output in the Genepop format (Rousset, 2008) which can be read by most population genetics data analysis programs like e.g. Arlequin (Excoffier et al., 2005). As ogaraK is, in fact, simulating an infinite population, the Genepop format (being based on a finite sample) is not enough to capture the genotype frequencies with full precision, therefore ogaraK also exports its results in a simple format with precise information of genotype frequencies over time. We here describe the parameters for ogaraK and the output format.

Example scenarios

Many example scenarios are available (please see the File/Open menu). They are named according to the following convention:

EpistasisMOI or
NumberofDrugsPolicyEpistasisMOI
Examples:

FullEpistasis2 contains an example for full epistasis with a MOI of 2.

Full+Assym4 is Full Epistasis and Asymmetry (SP-based) with MOI 4.

2MFTFull+DGF3 is 2 drugs, multiple first line therapies, Full Epistasis and DGF with MOI of 3.

2RotFull+DGF3 is 2 drugs, rotation, Full Epistasis and DGF with MOI of 3.

Parameters

ogaraK provides an easy to use interface. The following parameters are available:

Number of drugs The number of drugs used.

Drug type Either “Standard” or “Artemisinin”. In standard drugs, loci involved in drug resistance are independent from drug to drug. In Artemisinin drugs, there are always two loci per drug and one locus is shared among all drugs (see subsection A.2).

Loci per drug Number of loci encoding resistance per drug. Always two if the drug type is Artemisinin.

Epistasis Either “Full epistasis”, “Duplicate gene function” (DGF), “Full epistasis plus DGF” or “Full epistasis plus asymmetry” (see subsection A.1). For “Artemisinin” based drugs only “Full epistasis” is expected to be used, but the user can test alternative models.

Policy Either “Rotation”, “Multiple first line” or “combination”.

Rotation If policy is rotation, the user can specify the type of rotation: Either “Fixed”, i.e., every n generations, n being user specified or “Dynamic” where a drug is rotated after a certain resistance threshold for the drug being used is reached.

Infectivity Either “competitive release” or “independent transmission”. In competitive release an individual with 1 infection will produce as many secondary infections as one with 10. In independent transmission, the number of transmissions is proportional to the number of concurrent surviving infections.

Fitness penalty The type of penalty that will be computed, the most commonly used being “Erythrocytic”: Here any fitness penalty is modelled as competition in the blood stages, this means that the penalty is only useful in cases where the MOI is bigger than 1 (i.e., there is more than one infection competing in the blood). “Exo-erythrocytic” penalty exists mainly to model a penalty at the transmission and liver stages, i.e., it is MOI independent.

Penalty value The fitness penalty incurred per mutation (multiplicative). Specified as a range (see below).

Drug usage The percentage of infected humans that are treated. Untreated hosts function as sensitive reservoirs. Also specified as a range.

MOI The Multiplicities of Infection in the simulation. The user can specify from a MOI of 1 (selfing) to 7. All the MOIs must add up to 1. E.g. 0.25 of MOI 1 and 0.75 of MOI 2, means that 25% of all humans will have only one infection at a certain point in time, and the other 75% will have two.

Max generations The maximum generations for which each simulation can be run. A simulation can stop for two reasons: Either at the limit specified here or if the threshold of resistance is reached (see below).

Threshold The value of spread of resistance after which the simulation stops. It is normally not interesting to run the simulation above certain levels of resistance as after certain levels the drugs are not used anymore (as they have long lost efficacy).

Resistance introduction Activates mutation/immigration of resistants.

Prob. introduction The probability of introduction of a resistant per generation. Each locus will have the same probability specified here.

Percent replaced The percentage of sensitive alleles that will be replaced with a resistant version.

Initial frequency The initial frequency of the resistant forms.

Sample size The sample size for the Genepop file. While the model is not individual based, a sample of the every generation simulated can be produced. The sample size is approximated (it can be slightly lower than value requested) in order to try to assure a rounded approximation of the existing population.

Both fitness penalty and drug usage can be specified as a single value or as a range. If specified as a range, then more than one simulation will be run, example: Fitness is specified as range between 0 and 1 with a step of 0.1 and drug usage is specified as a range between 0.1 and 0.8 with a range of 0.05, this means that 11x15 simulations will be done, with fitness being 0.0, 0.1, ..., 1.0 and drug usage being 0.1, 0.15, ..., 0.8. If single values are specified for both parameters then only one simulation will be made.

It should be noted that the computational cost of running ogaraK simulations can vary widely. Users are recommended to read subsection A.4 before starting simulating with ogaraK.

Configuration has to be saved to a file before the simulator can be run. Configuration files can be easily edited by hand, this being especially useful for batch mode runs.

When running, the simulator writes all the simulations to disk in an internal format (in addition to output formats). This allows to re-use previous simulation results or to restart a batch of simulations that might be interrupted. If mutation is activated no re-use is done as results can vary due to stochasticity. All data files are written to the same directory where the configuration file is written.

Output

Three files are written for every simulation made: one with a sample of individuals in Genepop format – each population representing a different generation. A file with the frequency of all genotypes per generation and a file noting when drugs are rotated. The last file is only created when drug rotation policies are simulated. The files are

named with the following prefix: freqPOLICY-NUMDRUGSLOCIDRUG-PENALTY-DRUGUSAGE, e.g freqMFT-21-005-03 is a the prefix of files containing the results from an MFT policy with 2 drugs, 1 locus per drug, a fitness penalty of 0.05 and a drug usage of 0.3. The suffixes for file names are .txt for genepop files, .og for the frequencies and .sw for the swaps.

While the genepop file adheres to a widely used standard and thus needs no description, we provide here the description of the content of the frequency and swap files.

Frequency files

Each line has the frequency of a certain genotype configuration. This consists of two columns: the genotype and the frequency. All genotypes are enumerated. the process repeats for all generations simulated. Here is an example where a genome for 2 drugs and 1 loci per drug is simulated:

```
0 0.9
1 0.05
2 0.04
3 0.01
0 0.8
1 0.083
2 0.073
3 0.043
```

In this case 2 generations are shown: in the first the fully sensitive genotype has 90% of frequency, the one resistant to drug 1 has 5%, the one resistant to drug two has 4% and the multiple-resistant has 1%.

Drug swap files

Drug swap files depict when drugs are rotated, they format is very simple: each lines depicts when a drug is swapped, the first column indicates the generation and the second the drug introduced, example

```
0 0
7 1
14 0
```

Drug 1 (coded as 0) is introduced in generation 0, rotation first happens at generation 7, and again a 14.

A.4 Software issues

Complexity and computational cost

The number of different genotypes that must be tracked is equal to 2^l , where l is the number of drugs times the number of loci per drug (or the number of drugs plus 1 in the case of the Artemisinin model). As per the formula above all possible combinations of genotypes for each MOI will have to be considered. The number of permutations is then $2^{l^{MOI}}$. This makes the calculation above computationally very intensive, as an example studying a clonal multiplicity of 4 with 64 different genotypes (3 drugs with 2 loci per drug) requires considering 16 million combinations. The most extreme case theoretically allowed, with a MOI of 7 would require dealing 4,398,046,511,104 cases (which is not feasible in practice). Furthermore, this computation has to be done for each environment (untreated and each epistasis mode per drug), for every generation and for every simulation (the number of simulations being dependent on the ranges of both fitness penalty and drug usage).

Therefore, the user will have to decide on the best compromise between available computing resources and the parameter choice (noticing that different parameters have a completely different impact on the computational time).

OgaraK tries as much as possible to facilitate the run of computationally intensive simulations by doing the following:

1. A relative value of computational cost is presented to the user. Whenever the user changes parameters, an estimate of the relative cost can be made. While not being related to any physical measure of time, it gives a good relative comparison of time among different parameter choices.
2. Simulations are independent and can be run concurrently over a certain range of fitness and drug usage. OgaraK will generate different file names based on the parameters chosen, facilitating concurrent runs.
3. OgaraK can be run in batch mode (allowing multiple runs being starting automatically by a script on a single computer or on a grid), using a configuration file generated by the UI. This can be simple done by calling *java malaria.Article path_to_database path_to_config_file*.

Development infrastructure

The development infrastructure of ogaraK is fully open and public and relies on well established platforms for software development. Along with the user site <http://popgen.eu/soft/ogarak> there is also a development site <https://launchpad.net/ogarak> (the same platform used by projects like Ubuntu Linux or MySQL) which includes source

code development, bug tracking and test infrastructure among other facilities typical of professional development projects.

A.5 Example analysis

Here we present two very simple analysis of ogaraK results.

The first one is nothing more than plotting the change in frequency of all parasite types over time for a simulation (figure A.5). This example is presented only to exemplify the use of the frequency file (the code to generate the chart is available on the web site) and strictly as a starting point.

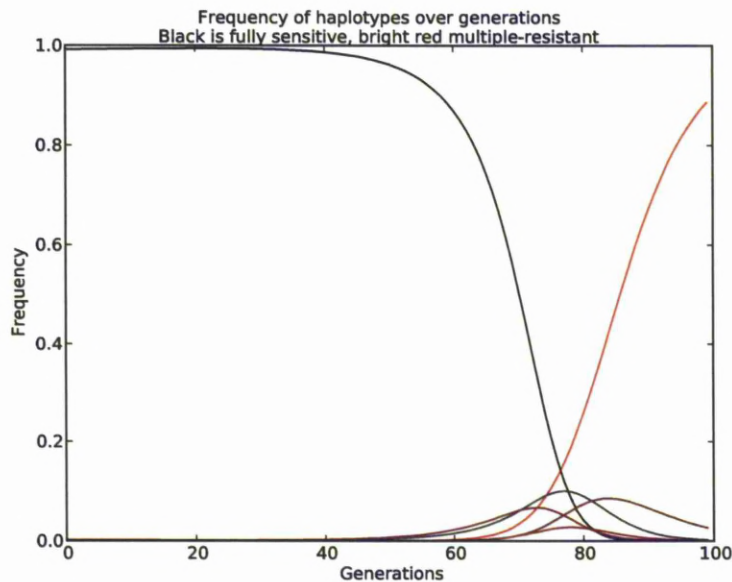
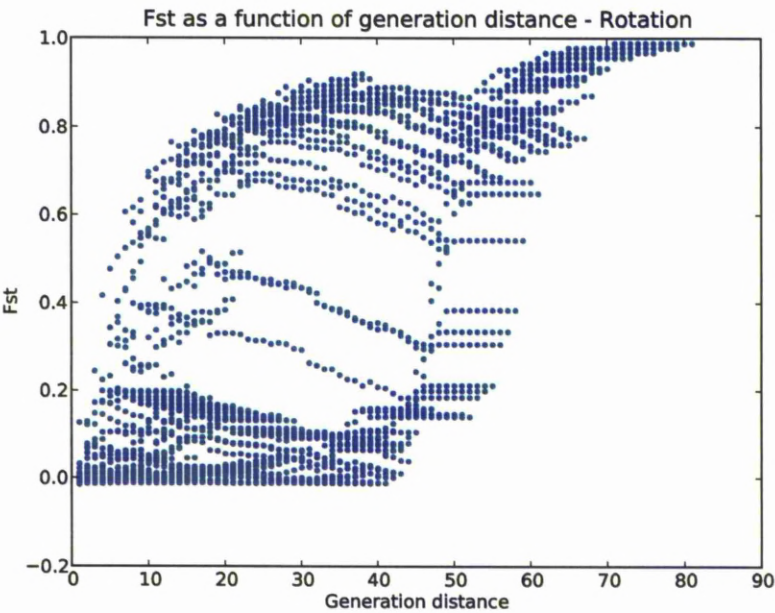


Figure A.5: The change in frequency of various genomes over time.

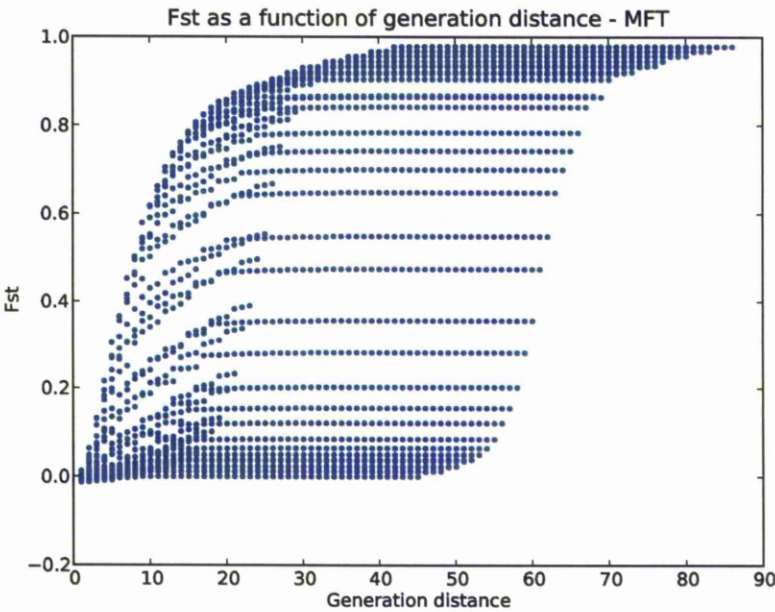
We expect that most data analysis made with ogaraK will be via the genepop output as genepop files can be feed to most population genetics analysis programs. The next example provided used the Biopython Genepop code to do the analysis. Two points should be noted: (i) coding skills are not required to do most analysis with genepop files – using one of the many applications will be enough and (ii) the example provided is one of the many that can be done, as the possibilities are unbound given the variety of methods available.

In our example, we compared the pattern of temporal F_{ST} when using rotation and MFT. On figure A.6 we plot the pairwise F_{ST} among all generations (between 0 and 100). The x-axis is the difference between the generations involved (e.g. if the comparison is between generation 7 and 20, this will fall on 13 on the x-axis). As expected F_{ST}

increases with distance in MFT as the selection pressure is the same over-time, but with rotation the pattern is less clear as the selection pressure is changing thus making the pressure for each individual drug intermittent.



(a) Rotation



(b) MFT

Figure A.6: How pairwise temporal F_{ST} varies with different treatment policies. In the x-axis the difference between compared generations is plotted.

B

A formal description of the population genetics model to study the spread of drug resistant malaria

Simulations were run varying drug usage (i.e., proportion of infections treated) and fitness penalty between 0.0 and 1.0 in increments of 0.02 for both parameters. The models were run with constant MOIs of 1 (100 % selfing), 2, 3 and 4. The application to perform these simulations is made available in <http://popgen.eu/soft/ogaraK>¹.

In order to facilitate understanding of the resistance profile of each clone we will use, whenever pertinent, a binary representation per locus where 0 represents a locus that is sensitive and 1 when a locus might confer resistance. As an example, in a case where we model two drugs and 1 resistance locus per drug, the totally sensitive clone can be represented by $c_{0,0}$; clone $c_{0,1}$ is sensitive to drug 1 and resistant to drug 2. For clarity and unless otherwise stated, all textual examples presented below will assume 2 drugs with 1 locus per drug.

The following symbols are also employed:

l is the total number of loci involved in drug resistance;

i is the concurrent number of infection on a human host or Multiplicity of infection (MOI);

F_k is the frequency of resistance profile k in the whole population in the current generation;

F'_k is the frequency resistance profile k in the whole population in the next generation;

f_k is the frequency of resistance profile k inside a single human individual in the current generation;

s is the strength of natural selection acting against each mutation;

m is the number of mutated loci in each clone, e.g. $c_{01,11}$ has $m = 3$;

¹ogaraK was accepted after this manuscript, therefore no reference was provided other than the web page.

d is the drug treatment rate, defined as the proportion of infected individuals treated. If more than one drug is used, then all are used in equal proportions.

n_d is the number of drugs used.

The calculations will be performed in a computer because, as will become apparent, the number of permutations of genotypes becomes huge. Here we provide the algebraic basis of the calculations; obviously we cannot describe each case so we will use illustrative examples.

Selection pressure is assumed to be mediated through competition between all clones in the asexual blood phase in humans. The proportion of transmissions contributed by a resistance profile k in any single host is:

$$t_k = \frac{f_k(1-s)^m}{\sum_{c=1}^i f_c(1-s)^{m_c}} \quad (\text{B.1})$$

When no drug is present; the denominator is the familiar “mean fitness” averaged across the c clones of the infection. If a drug is used, only infections that are drug resistant are able to transmit (i.e., $t_k = 0$ for sensitive forms). Being drug resistant depends on epistasis for models with more than 1 locus per drug: for full epistasis all loci are required to be resistant, for duplicate gene function a single locus is enough (and having more than one is slightly deleterious) and for asymmetrical epistasis, the first locus is required.

After successful transmission to a mosquito via a blood meal, *P. falciparum* reproduces sexually (with the possibility of selfing), and as such is subject to recombination. This means that new genotypes can originate inside the mosquito, as an example, if a clone resistant to drug 1 ($c_{0,1}$) mates with a clone resistant to drug 2 ($c_{1,0}$), 4 types of offspring are possible: the initial versions plus profiles $c_{0,0}$ and $c_{1,1}$, i.e., totally sensitive parasites and multidrug-resistant ones.

In the case of two drugs and one locus per drug and assuming that the recombination rate among loci is 0.5 (i.e., they are not physically linked) the frequency of transmission of sensitive parasites from a certain individual is given by:

$$f_{0,0} = t_{0,0}^2 + \frac{1}{2}t_{0,0}(t_{1,0} + t_{0,1}) + \frac{1}{4}t_{0,0}t_{1,1} + \frac{1}{4}t_{1,0}t_{0,1} \quad (\text{B.2})$$

A similar reasoning applies for all other resistance profiles.

In the general case, the probability of a certain infected human host being infected with MOI i composed of a certain combination of clonal genotypes is given by:

$$p(i)p(c_0, c_1, \dots, c_l) \quad (\text{B.3})$$

$p(c_0, c_1, \dots, c_l)$ is calculated from a multinomial distribution using current frequencies of clonal genotypes in the population.

In our example, this can be read as the probability of having i infections times the probability of having a combination of infections composed by $c_{0,0}$ sensitives, $c_{0,1}$ resistant to drug 1, $c_{1,0}$ resistant to drug 2 and $c_{1,1}$ multidrug-resistant.

The frequency of each resistance profile k transmitted to the next generation will be:

$$F'_k = \frac{\sum_{i=1}^i \sum_{c_0=0}^i \sum_{c_1=0}^{i-c_0} \dots \sum_{c_l=0}^{i-c_{l-1}-\dots-c_0} \sum_{e=0}^{n_d} \sum_{d=0}^{n_d} f(e)p(i)p(c_0, \dots, c_l)p(d)t_{e_k}}{\bar{W}} \quad (\text{B.4})$$

The structure of the equation can be explained as follows: summation over i allows the investigation of all MOI classes in the population (though in the examples researched, the value is fixed for the whole population, i.e., everyone has the same MOI). Summation over c_0, c_1, \dots, c_l allows inclusion of all possible combinations of genotypes within the MOI class. Summation of e allows to investigate the contribution of all environments (untreated individuals plus one or more different epistasis cases). $f(e)$ is the fraction of the host population providing a certain environment (e.g., untreated, treated with duplicate gene function, etc.). $p(d)$ is the probability of receiving a certain drug regimen where $d = 0$ means no drug. For the untreated environment $p(0) = 1$. For other environments $p(0) = 0$ and $p(d > 0)$ is dependent on drug policy (e.g. $\frac{1}{n_d}$ when using multiple therapies simultaneously in the same proportion). For each environment e there is a different transmission proportion for each profile (t_{e_k}).

\bar{W} is a normalization coefficient equal to the sum of all the numerators (in order to assure that the proportions of each type of transmission $\sum_{k=0}^{2^l-1} F'_k$ sum to 1).

A similar approach has been used before (Hastings, 2006), and is now extended to allow different environments, more than one drug and several models of epistasis. This more complex model has consequences on the computational cost of the simulation.

The number of different genotypes that must be tracked is equal to 2^l , where l is number of drugs times the number of loci per drug. As per the formula above all possible combinations of genotypes for each MOI have to be considered. The number of permutations is then 2^{li} . This makes the calculation above computationally very intensive. As an example studying a clonal multiplicity of 4 with 64 different genotypes (3 drugs with 2 loci per drug) requires considering 16 million cases. The most extreme case theoretically allowed, with a MOI of 7 would require dealing 4398046511104 cases (which is not feasible in practice). Furthermore, this computation has to be done for each environment (untreated and each epistasis mode per drug), for every generation and for every simulation (the number of simulations being dependent on the ranges of both fitness penalty and drug usage).

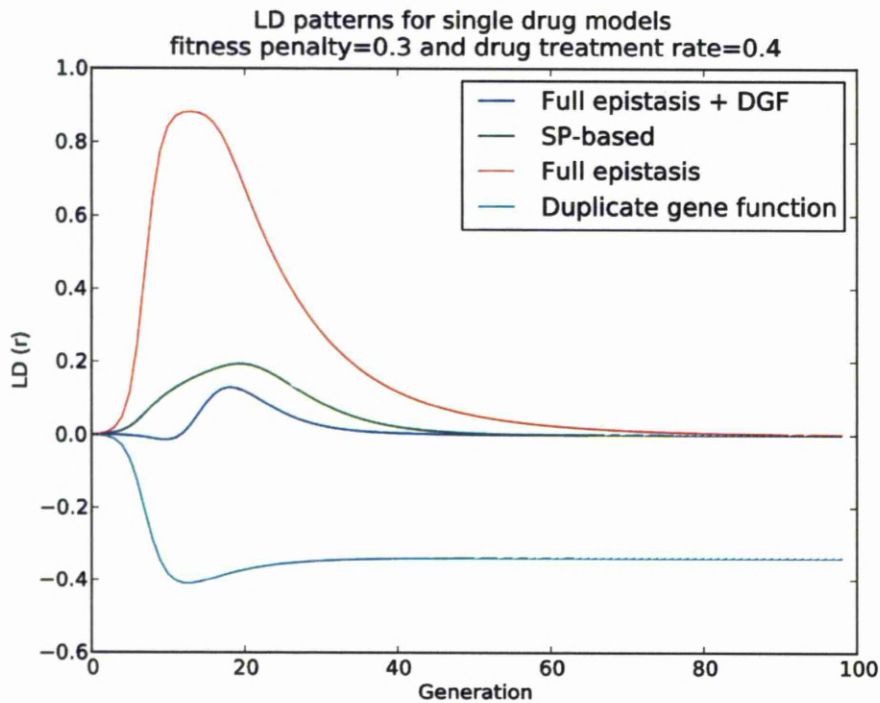
Linkage (gametic) disequilibrium is expected to be influenced by epistasis and is a necessary parameter to understand the relationship between frequency and prevalence of resistance. In order to understand and compare the impact of epistasis on linkage, we studied linkage disequilibrium using the correlation coefficient r which, while not

completely removing the effects of allele frequency (Hedrick, 1987), still allows for a standardised qualitative comparison required for this study. Furthermore, unlike D' , it preserves the original signal of D which is important to understand epistasis effects.

Epidemiologically realistic MOIs

While the main manuscript only discusses scenarios with constant MOIs (i.e. all humans have the same number of infections), epidemiologically realistic settings have mixed MOIs (Arnot, 1998). The model presented is capable of simulating mixed MOIs but, from a qualitative point of view the results are similar to constant MOIs. Therefore, for the sake of simplicity, the main text manuscript presents results with constant MOIs.

Below we present a chart of LD patterns using a more realistic truncated Poisson distribution of MOI (truncated at 7 clones with a conditional mean of 2.5). As it can be observed, the chart is qualitatively similar to Figure 2 on the main manuscript (constant MOI of 2).



C

Expected heterozygosity: Illustrative examples of a slow moving statistic

Here we illustrate, using a computational study, the behaviour of a widely used long-term N_e estimator based on expected heterozygosity. The simulation results illustrate that the long-term estimator, and heterozygosity in general, cannot be used to detect the effectiveness of control and elimination interventions.

We also try to make suggestions about sampling sizes required to obtain a reliable estimation of heterozygosity based N_e estimation: How many individual samples and number of loci are needed for an accurate and precise N_e estimation with both estimators?

While we produce examples with scenarios based on effective population size of *P. falciparum* malaria, but our conclusions are generally applicable to all species where N_e is high (i.e. above 1,000). These illustrative simulations are thus concerned with the impact of demographic events and sampling strategies on heterozygosity and N_e estimation and interpretation. Other important factors, like the assumptions on mutation model and rate are not discussed here as they were subject of previous work.

This supplement follows roughly the structure of a research paper, but the results presented are not novel, only illustrative of known properties of expected heterozygosity. Readers might want to skip the methodological details and read directly the results and conclusions or even just inspect the figures and tables.

C.1 Methods

The long-term N_e estimator based on heterozygosity assumes that in a population in mutation-drift equilibrium heterozygosity is a function of the product of N_e and the mutation rate μ . Assuming a stepwise mutation model (SMM) expected in many microsatellite markers the relationship between N_e , the mutation rate μ and H_e :

$$N_e\mu = \frac{\frac{1}{1-H_e} - 1}{8} \quad (\text{C.1})$$

We also studied the infinite alleles model (IAM), used in loci with patterns of variation incompatible with the SMM (Anderson et al., 2000a), but results are not shown as they are qualitatively similar to SMM results.

Expected heterozygosity (H) at each locus is calculated as $\frac{n}{n-1}(1 - \sum_{i=1}^n p_i^2)$, where n is the number of markers sampled and p_i is the frequency of the i th allele and the factor $\frac{n}{(n-1)}$ is a correction to allow comparison of results obtained from loci with different numbers of samples.

Simulations

We conducted simulations using the forward-time, individual based simulator simuPOP (Peng and Kimmel, 2005). Simulations were done using a diploid genome with 100 neutral independent microsatellites, random mating, discrete generations, random variation in reproductive success and an average sex ratio of 1. Initial allele frequencies were constructed in order to replicate the allele frequency distribution expected (both in H_e and number of alleles) for the Stepwise Mutation Model estimating an N_e as observed in empirical studies (Iwagami et al., 2009; Anderson et al., 2000a; Susomboon et al., 2008) using a microsatellite mutation rate is estimated as 1.59×10^{-4} for *P. falciparum* malaria (Anderson et al., 2000b). The number of initial alleles was varied between 5 and 24 as observed in Anderson et al. (2000a). For all simulations 100 replicates were conducted. Haploid simulations were also conducted in order to research the impact of ploidy as *P. falciparum* has both a haploid and a diploid phase.

To quantify the impact on bias, precision and accuracy of sample sizes, we simulated 5 constant size populations with N_c (census size) of 20,000, 10,000, 5,000, 1,000 and 500. As the simulated populations are assumed to be panmictic and of constant size, N_c is an approximation of the contemporary N_e (Charlesworth, 2009). The sampling strategies include 10, 20, 50 90, 200 and 300 individuals using 5, 10, 20, 100, 200 and 500 loci.

The ability to detect a sudden contemporary bottleneck (a consequence of a control or elimination intervention) was simulated using population of initial N_e of 20,000, 10,000, 5,000 and 1,000 (based on values determined for *P. falciparum* population in high, medium and low transmission sites (Anderson et al., 2000a; Susomboon et al., 2008)) bottlenecked to 50%, 10% and 5% of its original size. We also simulated expansions varying between 2 and 20 times the original size. We then sampled the population several generations after the bottleneck/expansion and estimated the N_e using the heterozygosity and LD methods. The N_e before the demographic event is termed N_1 and the N_e afterwards is N_2 .

To estimate the importance of founding effects we also simulated an initial N_e as per the paragraph above then followed by 2 to 10 generations where N_e is either 10

or 100 (a severe bottleneck representing a minority of individuals migrating to found a new population, or after a very severe and successful control programme) and then a population expansion to 1,000, 10,000 and 20,000.

C.2 Results and remarks

In low transmission scenarios (N_c below 2,000) the precision of the estimator is moderately dependent on the number of loci and individuals sampled. With a strategy containing around 30 haploid samples and 10 loci (common in studies for low transmission) half of estimates are greater than 20% from the estimated median value. Similar precision patterns are observed for high transmission scenarios ($N_e > 10,000$). Strong upward bias is observed with all sampling strategies especially in high transmission simulations. Bias is stronger with smaller samples (the median estimate is around 15% above the real N_e) whereas bias with the larger sampling strategy is only around 5%. Bias is higher when the sample size approaches the number of alleles as this causes the allele distribution to be more uniform (therefore increasing heterozygosity). We present the distribution of point estimates using multiple sample strategies for a constant N_e of 1,000 and 20,000 on figure C.1.

In order to understand the relationship between of number of loci with the number of individuals genotyped, we used two sampling strategies where the number of loci and individuals is swapped (i.e., one with 10 loci and 100 haploid individuals and another with 100 loci and 10 individuals). Accuracy is similar in all scenarios (constant, bottleneck and expansion) tested. The two sampling strategies are compared on figure C.1 for constant N_c simulations (last two columns). Most importantly interchangeability of number of loci and individuals also holds for bottleneck and expansion simulations. Figure C.2 shows \hat{N}_e for various bottlenecks and expansions using both sampling strategies.

In bottleneck and expansion scenarios, estimated N_e values overwhelmingly reflect pre-bottleneck N_e values even after many generations. This happens for all values of N_e (pre- and post-bottleneck/expansion) typically reported for *P. falciparum*. Even on the most extreme scenario, modeling a population decrease from a value typical of high-transmission settings (N_e of 20,000) to a value typical low-transmission value (1,000), after 100 generations of the bottleneck the estimated N_e is still, on average, 75% of the original value or, around 15 times the post-bottleneck value. For a decrease from 20,000 to 5,000, no estimation decrease whatsoever (due to upward bias) can be detected 100 generations after the bottleneck. The effect is worse with expansion scenarios: even in the most extreme case of an expansion from 1,000 to 20,000, after 100 generations the estimator is only 10% above the original value. Bottlenecks are detected faster as heterozygosity is decreased via drift whereas in expansion heterozygosity is increased via mutation, a slower process. Table C.1 shows mean H_e and \hat{N}_e values for several

generations after a range of bottlenecks and expansions using a sampling strategy of 200 diploid individuals and 100 loci.

With founding models, any estimated N_e is mostly influenced by pre-bottleneck size and the size and duration of the migrated population. Contemporary population size has minor impact on the estimated value. If the founding population is extremely small or the time to found (i.e., the number of generations between leaving the original population and starting to expand the new population) is large, then the estimated N_e post-founding will be strongly influenced by the migrated population. For instance when only 10 individuals migrate, and assuming that migration takes 2 generations, there is an average loss of heterozygosity of 0.1 (consistent with a theoretical expectation of heterozygosity loss of $\frac{1}{2N}$ per generation) causing a proportional fall in N_e . If the migration is either of a large enough population or of small duration then the N_e estimation post-bottleneck will mostly approximate the value pre-migration. Figure C.3 shows \hat{N}_e up to 20 generations post-founding assuming a pre-migration N_e of 20,000 and stable post-founding N_e of 1,000, contemporary N_e mostly influences the slope which is never steep for typical *P. falciparum* values.

As long as the sample size is enough to accommodate the number of alleles, the number of individuals sampled can be exchanged for more loci. This means that, if for some reason it is not possible to sample many individuals then sampling more loci per individual can compensate in order to increase the accuracy. This is especially important because in some situations it is not possible to genotype more individuals but with next-generation sequencing, it is feasible to genotype many loci per parasite. The ability to exchange individual numbers for loci number happens in constant, bottleneck and expansion scenarios. Evaluating the accuracy trade-off of the estimator is particularly important in bottleneck scenarios because it has been demonstrated that, with contemporary N_e estimators, sampling more individuals is more informative than sampling more loci (England et al., 2010; Antao, 2010).

The bias and accuracy issues that we researched are due to demographic events and sampling strategies. These effects will have to be compounded with the already well known impact of varying mutation rates. Indeed, our models took a very benevolent approach to both mutation rate and mutation model as we assumed a constant mutation rate for all loci based on the value commonly used for N_e estimations with *P. falciparum* malaria and a Stepwise Mutation Model (SMM) whereas some loci show patterns of variation which are inconsistent with the SMM (Anderson et al., 2000b). Even with such assumptions, several limitations of this estimator of effective population size became clear.

N_1	N_2		Generation					
			1	5	10	25	50	100
20000	10000	H_e	0.811	0.811	0.811	0.811	0.811	0.811
		\hat{N}_e	21196	21203	21201	21260	21151	21138
20000	5000	H_e	0.811	0.811	0.810	0.810	0.809	0.806
		\hat{N}_e	21238	21152	21101	20936	20698	20184
20000	1000	H_e	0.811	0.809	0.807	0.802	0.793	0.776
		\hat{N}_e	21159	20721	20289	19178	17510	14924
10000	20000	H_e	0.730	0.730	0.731	0.731	0.732	0.733
		\hat{N}_e	10018	10033	10041	10073	10134	10226
10000	5000	H_e	0.731	0.731	0.731	0.730	0.729	0.728
		\hat{N}_e	10078	10044	10044	10018	9952	9814
10000	1000	H_e	0.730	0.728	0.727	0.722	0.714	0.699
		\hat{N}_e	9991	9849	9731	9408	8834	7903
5000	20000	H_e	0.630	0.630	0.630	0.631	0.632	0.634
		\hat{N}_e	4946	4956	4969	4986	5018	5093
5000	10000	H_e	0.630	0.630	0.630	0.631	0.632	0.634
		\hat{N}_e	4956	4960	4964	4979	5007	5067
5000	1000	H_e	0.630	0.629	0.628	0.624	0.618	0.607
		\hat{N}_e	4949	4920	4884	4780	4601	4296
1000	20000	H_e	0.340	0.342	0.341	0.344	0.349	0.357
		\hat{N}_e	1017	1027	1025	1041	1067	1113
1000	10000	H_e	0.340	0.341	0.342	0.344	0.349	0.356
		\hat{N}_e	1020	1023	1027	1042	1066	1112
1000	5000	H_e	0.340	0.341	0.343	0.344	0.348	0.355
		\hat{N}_e	1020	1025	1033	1043	1061	1101

Table C.1: Mean heterozygosity and effective population size estimated after a bottleneck or expansion. N_1 is the value before the bottleneck/expansion and N_2 the value after. Estimations are done 1, 5, 10, 25 50 and 100 generations after the demographic event. The sampling strategy includes 200 diploid individuals and 100 loci.

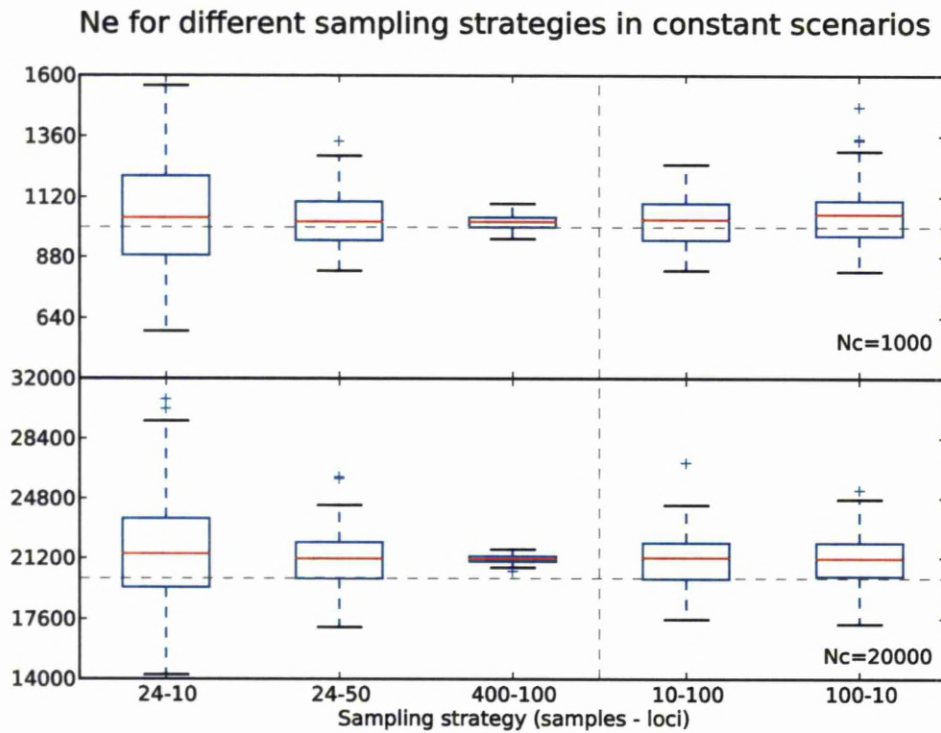


Figure C.1: Boxplot charts of the distribution of N_e point estimates for a scenario of constant N_e of 1,000 (top chart) and 20,000 (bottom chart) using different sampling strategies. The sampling policy is specified on the X-axis: the first number is the number of samples and the second number, the number of loci. Sampling policies vary from 48 diploid samples with 10 loci, typical of empirical studies in malaria low-transmission areas to 200 diploid individuals and 100 loci. The last two columns compare the behaviour of the estimator with sampling strategies exchanging loci for individuals.

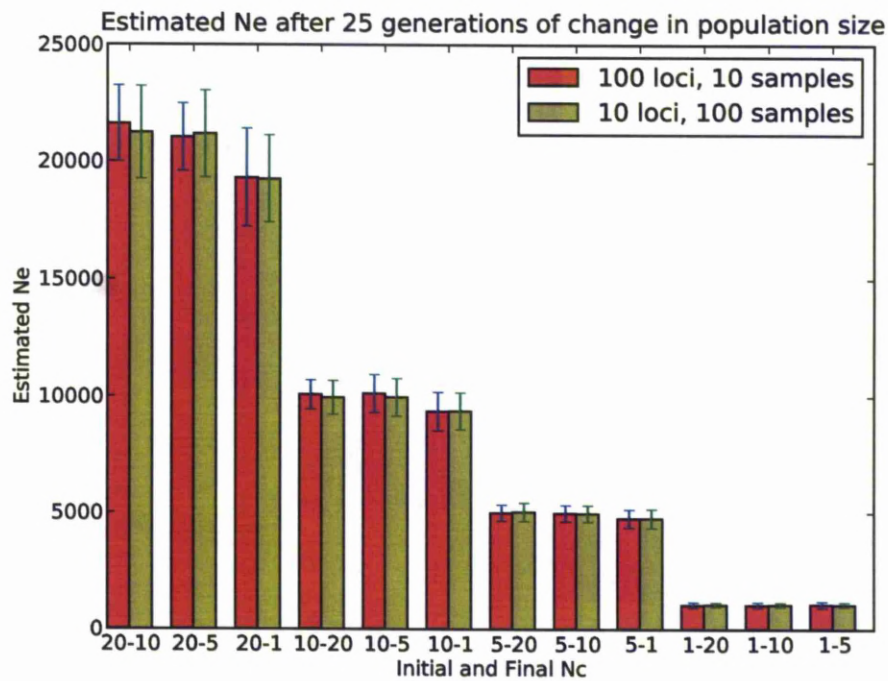


Figure C.2: Bar charts for the estimation of N_e in bottleneck and expansion scenarios where sampling occurs 25 generations after the bottleneck/expansion. Two sampling strategies are used one maximising the number of loci (100 loci, 10 individuals), the other the number of individuals (10 loci, 100 individuals). The Y-axis is the estimated N_e and the X-axis includes both N_c before and after the bottleneck/expansion event. Mean and standard deviation are plotted.

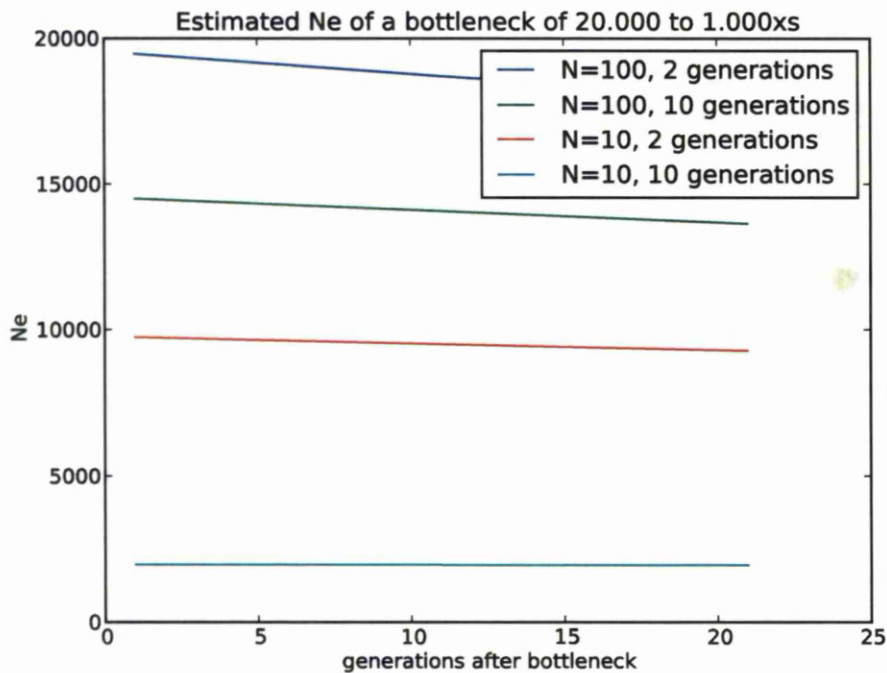


Figure C.3: Mean \hat{N}_e for four founding scenarios from an initial population of 20,000 to a final population of 1,000 using 200 diploid individuals and 100 loci. The founding population has 10 or 100 individuals and during 2 or 10 generations. The estimated value is mostly a function of the original N_e and the size and duration of the founding effect. Notably the influence of the contemporary N_e is minimal and only causes a slight slope over time.